IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE EDUCATION SOCIETY SECTION

Received 6 March 2026, accepted 21 March 2026, date of publication 30 March 2026, date of current version 2 April 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3678846

SURVEY

Knowledge Tracing Models in Educational Data Mining: Historical Evolution, Categorization, and Empirical Evaluation

PRINCE DAS ADHIKARY[1], JARI METSÄMUURONEN[1], MIKKO-JUSSI LAAKSO[1], AND JUKKA HEIKKONEN[2]

[1]Turku Research Institute for Learning Analytics (TRILA), University of Turku, 20014 Turku, Finland
[2]Department of Information Technology, University of Turku, 20014 Turku, Finland

Corresponding author: Prince Das Adhikary (pdaadh@utu.fi)

This work was supported in part by the Education for the Future (EDUCA) Flagship through the Research Council of Finland under Grant 358924 and Grant 358947, and in part by the EDUCA-Doc Doctoral Education Pilot through the Ministry of Education and Culture (Doctoral School Pilot) under Grant VN/3137/2024-OKM-4.

ABSTRACT This article analyses computational models of Knowledge Tracing (KT), which address the complex sequence-modelling task of predicting dynamic, unobservable latent user states from historical interaction logs. First, we propose a comprehensive taxonomy identifying nine distinct and interconnected KT model families: psychometric; Bayesian; machine learning; deep learning; graph-based; temporal/sequential; multi-task; contrastive/self-supervised; and domain-adaptive. Secondly, we trace the historical evolution of KT architectures, from the foundational psychometric methods of the 1950s to the modern integration of attention mechanisms and graph neural networks. Thirdly, we systematically evaluate nine lightweight representative computational models—one from each category—across two large-scale datasets: ASSISTments 09-10 and DigiArvi 2025. We measure predictive calibration using accuracy, F1 score, ROC-AUC, average precision, and log loss under a strict computational time budget. Finally, our rigorous empirical analysis demonstrates that multi-task and temporal/sequential architectures yield the highest performance. Specifically, Fine-Grained Knowledge Tracing (FKT) achieved the best results on the DigiArvi dataset (accuracy: 0.77; F1 score: 0.85), while Temporal Item Response Theory (TIRT) performed best on the ASSISTments dataset (accuracy: 0.70; F1 score: 0.75). Traditional baselines, such as Logistic Regression (LR), remain highly competitive. Consequently, we advocate a shift towards 'Green AI' and standardized benchmarking to address the field's fragmented evaluation standards, as we identify diminishing returns from increasing model complexity. Future research must leverage generative Artificial Intelligence (AI) and causal inference to move beyond simple prediction toward agentic AI systems capable of active pedagogical intervention.

INDEX TERMS Knowledge tracing, deep learning, machine learning, psychometric, attention mechanism, Bayesian.

I. INTRODUCTION

Knowledge Tracing (KT) is a foundational task in Educational Data Mining (EDM). It involves tracing and predicting a learner's knowledge state over time by evaluating the acquisition of Knowledge Components (KCs). By treating learner-system engagement as a sequential data problem, KT analyzes historical interaction logs to optimize Intelligent Tutoring Systems (ITS) and algorithmic learning frameworks [1], [2]. This article presents a systematic review complemented by an empirical analysis of representative KT models from each major category, offering valuable insights for researchers (see Section III-C for an analysis of the selected models in each category).

The associate editor coordinating the review of this manuscript and approving it for publication was Nkaepe Olaniyi.

Recent advancements in KT have focused on improving prediction accuracy and overcoming the limitations of existing models. Modern KT models use machine learning and deep learning to predict student performance from historical data [3], [4], [5]. These models also leverage neural networks to capture patterns [6], [7]. In addition to deep learning models, many advanced KT models, such as multi-task models, use auxiliary tasks to support the primary prediction [8]. Domain adaptation models excel at transfer learning, in which models are fine-tuned on earlier pretraining data rather than trained from scratch [9]. Traditional KT models include Bayesian models, which are foundational probabilistic models [10], and psychometric models, which often use item response theory to model student performance [11].

This article examines various KT models, their historical development, and their interconnections. This approach demonstrates the evolution of KT models from traditional to modern methods. The primary goal is to provide a comprehensive overview of KT models, illustrating their historical and contemporary interconnections and anticipating future trends. The study presents the history of KT as a timeline, showing how models are built on past ideas while focusing on the latest trends. Additionally, a quantitative evaluation of representative KT models from each category was conducted on two datasets: ASSISTments 09-10 and DigiArvi 2025. The analysis focuses on how different categories perform across common metrics.

A. RESEARCH QUESTIONS

The four primary research questions explored in this study are the following:

RQ1: What is the big picture for the different KT categories in educational data mining, and how are they interconnected?

RQ2: How did KT models evolve up to 2025?

RQ3: What are the key categories of KT models in educational data mining, and how do representative models from each category perform across datasets?

RQ4: How do representative KT models compare across evaluation metrics, datasets, prediction curves, and student-level predictions?

To address RQ1, RQ2, and parts of RQ3 (specifically for category formulation), a systematic literature review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [12] guidelines. Additionally, an empirical analysis was performed to answer RQ3 (partly) and RQ4.

B. CONTRIBUTION

This survey paper aims to provide an extensive collection of KT architectures and clarify their boundaries for a wider audience of researchers, developers, and educators. Key contributions of this survey include:

- A systematic review of the main elements that comprise KT models, identifying nine distinct and interconnected model families (including psychometric, deep learning, and graph-based approaches), emphasizing how these systems have evolved from the 1950s to 2025.
- A comprehensive study of the techniques and concepts employed in the construction and evaluation of KT models, featuring a detailed timeline of architectural shifts from basic probabilistic methods to contemporary self-supervised and attention-based mechanisms.
- A rigorous empirical assessment of lightweight representative models across distinct educational datasets (ASSISTments 09–10 and DigiArvi 2025) under a strict computational time budget, providing practical examples of the effectiveness and limitations of different KT models in real-world scenarios.
- A critical discussion on the engineering trade-offs between predictive accuracy and model complexity, highlighting the diminishing returns of vast deep learning networks and demonstrating the specific advantages of specialized Temporal Item Response Theory (TIRT) and multi-task Fine-Grained Knowledge Tracing (FKT) models.
- The advocacy of a standardized benchmarking framework and 'Green AI' principles to address the fragmented evaluation standards within the field, ensuring that future systems are accurate, computationally sustainable, and reproducible.
- Actionable recommendations for future research detailing how Large Language Models (LLMs), causal inference, and efficiency metrics can be integrated to evolve KT from simple prediction algorithms to agentic Artificial Intelligence (AI) systems capable of active pedagogical intervention.

The expected contributions of this paper go beyond a mere literature review, instead providing a useful, empirically backed, and well-organized background to the issues and specifics of modern knowledge tracing. The remainder of the paper is structured as follows: Section II details the PRISMA systematic review methodology and the empirical evaluation setup. Section III provides an overview of KT categories, a timeline analysis of their evolution, an examination of specific model families, and a detailed empirical analysis. Section IV discusses the study's limitations, actionable future research directions, and practical recommendations. Finally, Section V concludes the study.

II. METHODOLOGY

A. SYSTEMATIC REVIEW PROCESS

This study employs a systematic review methodology that adheres to the PRISMA 2020 guidelines [12]. Applying the PRISMA structure ensures a transparent, exhaustive, and reproducible process for searching, selecting, and evaluating relevant KT models for education technology. Using the guidelines ensures that the literature review is methodical and systematic, including high-caliber work and avoiding the risk of bias.

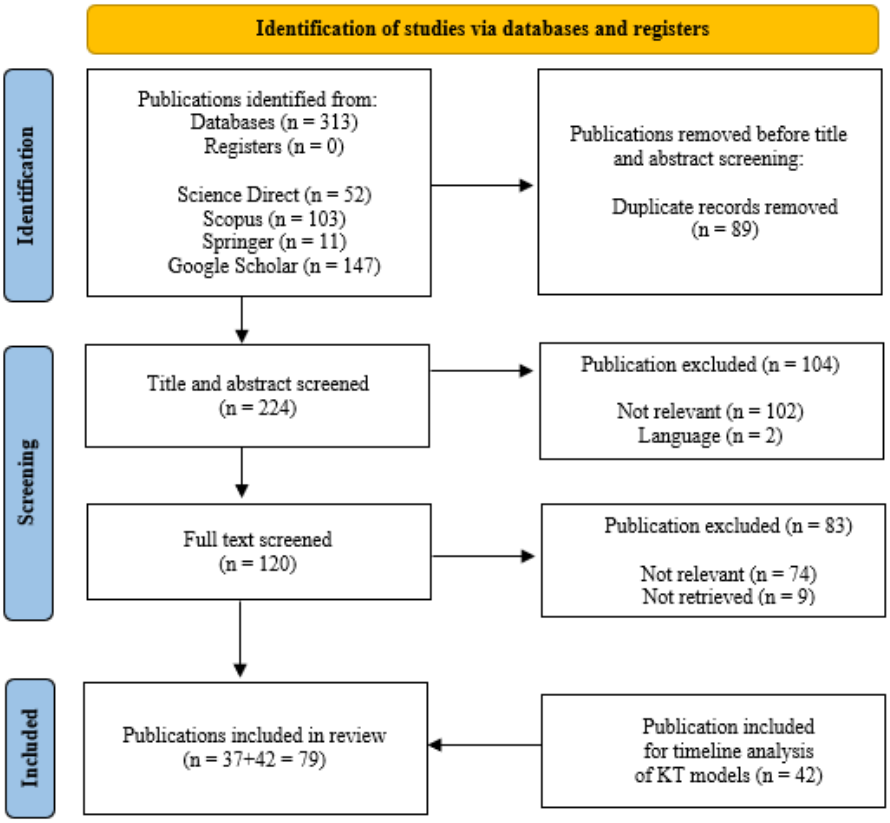

FIGURE 1. PRISMA flowchart for screening and identification.

A comprehensive search was conducted using central academic databases (Springer, ScienceDirect, Scopus, and Google Scholar). The search sought papers from 2019 to 2025. To ensure the collection is relevant and focused, the following search terms were used:

("digital trace" OR "digital footprint") AND ("artificial intelligence" OR "machine learning") AND ("learner" OR "student") AND ("skill assessment" OR "learning outcome")

The search strategy used Boolean operators (e.g., AND, OR) with keywords to restrict the search to studies comparing KT models for online education platforms, including those used by universities, K-12 schools, and colleges.

Inclusion and exclusion criteria were applied to methodically screen the studies. Inclusion criteria included peer-reviewed journal papers, conference papers, papers from the years 2019 to 2025, papers proposing, evaluating, or comparing KT models, and papers highlighting AI-based models for education. Exclusion criteria included non-English papers, duplicate papers, and papers unavailable for screening. This process filtered out irrelevant, low-quality papers.

Following the PRISMA guidelines, the study adopted a multi-stage screening and selection process. Initially, the database search yielded 313 papers. After excluding 89 duplicate records, screening by title and abstract identified 224 papers. This filtered out 104 papers as non-applicable (102) or non-English (2). Then, the remaining 120 papers were assessed based on their full texts. The methodological rigor, applicability of the subject matter to the field of KT, and empirical assessment of the models were evaluated. This process eliminated 83 papers, including 74 that were irrelevant to the subject matter and nine that could not be retrieved, leaving 37 papers. Additionally, to support the development of the timeline of KT models, an additional 42 studies were explored. Thus, a total of 79 papers were reviewed in detail to inform the systematic review and timeline analysis. Additional references were cited throughout the manuscript to provide the necessary background information on evaluation metrics, 'Green AI' principles, and the broader context of the discussion.

Fig. 1 shows the PRISMA flow diagram depicting how the studies were selected, step by step, from the initial search to the final 37 studies. The flowchart clarifies the selection process and ensures that the study was conducted clearly and fairly.

A relevant literature review was conducted to explore RQ1 and RQ2, as well as parts of RQ3 (specifically, category formulation). KT categories were developed based on

TABLE 1. Sample data for training nine representative KT models.

IDCode	orig_order	item	group	response	response_time_sec	sex
21432	1	71516	choose_1	1	25.167	M
21825	1	48881	algebra	0	21.942	F
21825	2	48882	choose_1	1	7.38	F
21825	3	48883	algebra	0	5.509	F
21825	4	48884	algebra	1	12.003	F
…	…	…	…	…	…	…

significance and current trends in the field rather than existing literature categories. This process identified nine emerging categories (see Section III-C). Additionally, the development of KT models was examined through a literature review, which provided a detailed overview of their progression over time (see Section III-B). Although there was some overlap between the literature on KT categories (systematic review) and the additional literature on timeline progression, the interconnections between different KT categories were examined by analyzing the models within each category to understand how they relate to one another.

B. EMPIRICAL EVALUATION SETUP

Two educational datasets were used in the empirical analysis to provide context for the literature review. The first dataset, ASSISTments 09–10, is derived from an online grade-school mathematics tutor [13]. The second dataset, DigiArvi 2025 (also known as DigiEva 2025), comprises computer-based math assessments administered to Finnish students in grades 3, 6, 8, and 9 [14].

Table 1 shows the sample data used to train the KT models for each category (see Section III-C) to predict the probability that a student will answer the next item correctly. For training purposes, the raw data has been transformed into the specified columns. To evaluate model performance, the study uses five complementary metrics. Accuracy counts the proportion of correct predictions; the F1 score balances precision and recall via its harmonic mean; the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) measures a model's ability to distinguish classes across all decision thresholds by taking the area under the ROC curve. Average Precision (AP) summarizes the Precision-Recall (PR) curve as the area under it [15]. Finally, log loss (cross-entropy) gauges probabilistic calibration by penalizing confident yet incorrect predictions [16].

To ensure data quality and reproducibility, a standardized pre-processing pipeline was applied to both datasets. Students with fewer than five interactions were filtered out to reduce noise from sparse data sequences. Variable-length sequence padding was also used to preserve complete learning trajectories. For model training and evaluation, a user-stratified 80/20 train-test split was used with a fixed random seed (random_state = 42). This chronological split ensures that 80% of each student's sequential interactions are used for training, with the remaining 20% reserved for testing. This rigorously evaluates the models' ability to generalize to new data without revealing future interactions. To ensure a rigorous evaluation, all nine representative models were implemented in Python, with all neural architectures constructed in PyTorch to ensure consistency. In line with our 'Green AI' principles [17], we deliberately selected lightweight, foundational architectures from each family (e.g., standard Deep Knowledge Tracing (DKT) rather than heavily parameterized transformer models). We tailored our optimization strategies to the architectural families: classical baselines used standard, appropriate solvers (e.g., maximum likelihood estimation for Bayesian Knowledge Tracing (BKT) and SAGA solvers for Logistic Regression (LR) and TIRT), while deep learning architectures were optimized using the Adam optimizer and categorical cross-entropy. Finally, to strictly evaluate algorithmic sustainability, we enforced a global training time budget of 120 seconds per model. To ensure transparent reporting of these computational constraints, all experiments were conducted locally on a machine with an Intel Core i3 processor and 16 GB of RAM, and no dedicated GPU was used.

III. RESULTS

A. THE BIG PICTURE OF KT CATEGORIES (RQ1)

This section provides an overview of the key KT categories and their intersections, offering insight into how these families collectively influence the field of modern EDM. Fig. 2 presents a Venn diagram illustrating the relationships among the nine distinct KT model categories, which are detailed further in Section III-C. As shown, the overarching category is machine learning, which encompasses several prominent subfields, most notably deep learning. Specialized approaches, such as graph-based and multi-task models, are deeply embedded within the deep learning paradigm. Together, they can leverage student interactions within a graph structure to predict future learning behaviors and mastery levels. Domain-adaptive, multi-task, and contrastive/self-supervised learning enhance deep learning's capabilities with transfer learning, auxiliary tasks that improve core predictions, and the ability to predict with scarce or unlabeled data. These overlapping areas demonstrate that KT is not confined to a single category but integrates multiple categories to enhance the model's effectiveness. The intersection of

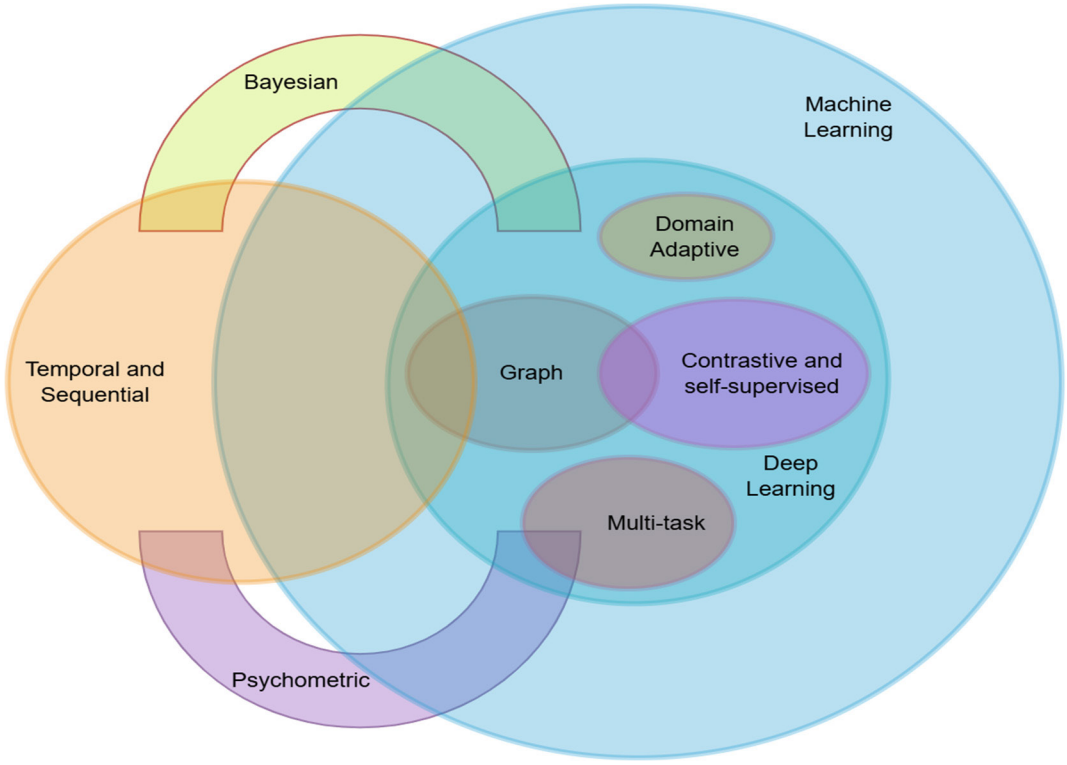

FIGURE 2. Venn diagram for categories of KT models.

Bayesian and temporal/sequential categories illustrates how learners' knowledge evolves as their latent states change over time. Similarly, the relationship between psychometric and temporal/sequential categories accounts for the integration of time into psychometric principles.

The diagram illustrates the increasing trend of personalized learning, in which models are calibrated to meet the unique needs of individuals in educational settings. In this case, we can see that advancements in deep learning models are being made to a reasonable extent in the current

KT scenario. This paves the way for exploring how various methods work together to improve student outcomes and, consequently, enhance learning analytics in the modern era. To categorize these models systematically, particularly when their architectures overlap across multiple domains (as illustrated in Fig. 2), we established classification boundaries based on the core algorithmic innovation or primary objective of each model. For instance, if a model's fundamental structure is based on neural networks, but its main contribution is introducing an additional task, it is classified under the 'multi-task' category rather than the broader 'deep learning' category. Similarly, models that integrate time into traditional psychometric or statistical structures are classified under 'temporal and sequential' if their defining feature is the explicit modelling of temporal dynamics.

B. TIMELINE DEVELOPMENT OF KT MODELS (RQ2)

To understand the current state of KT, it is important to trace its historical development. From a timeline perspective, the development of KT architectures can be categorized into four distinct eras of progressive development: the Foundations of KT (1950–1989); Early Innovations and Refinements (1990–2014); the Deep Learning Surge (2015–2019); and Next-Gen Innovations (2020–2025).

1) FOUNDATIONS OF KT (1950–1989)

The evolution of KT began in 1950 and continued until 1989, when core fundamental theories began to emerge. During this period, there were many changes, from the initial psychometric, statistical, and probabilistic models to the early stages of machine learning. These early stages of machine learning eventually became the backbone of KT models.

The initial psychometric models used early Item Response Theory (IRT) to show the relationship between learners' latent abilities and their probability of answering test items correctly. Pioneers of this work include Rasch [18] and Lord, Novick, and Birnbaum [19]. Multidimensional Item Response Theory (MIRT) models multiple latent traits, providing a more comprehensive view of individual traits [20].

Regarding the early statistical and probabilistic models, LR was used to understand the probability of correct responses in an evaluation [21]. A Naïve Bayes (NB) classifier, which uses Bayes's theorem for feature independence, was also employed in the analysis [22].

Early machine learning models, such as the Decision Tree (DT) algorithm, are used to make structured decisions based on classification [23]. Additionally, Bayesian networks (BNs) provide a graphical representa-

tion of the probabilistic relationships between variables [24]. K-means clustering divides data into groups based on pattern similarity [25].

From 1950 to 1989, core KT models were formulated that paved the way for the far more advanced architectures of today. One example is IRT, which describes the connection between latent ability and the probability of providing the correct answer. Then, MIRT describes the relationship between latent ability and multiple traits. Early probabilistic and statistical models, such as LR and NB classifiers, alongside early machine learning models like DT, BN, and K-means clustering, pioneered the structure and data-driven analysis of learning. These advancements laid the groundwork for future models.

2) EARLY INNOVATIONS AND REFINEMENTS OF KT (1990–2014)

Between 1990 and 2014, core Bayesian models emerged, ideally paving the way for dynamic learner tracking. The Vanilla Bayesian Knowledge Tracing model, also known as BKT, provided a framework to analyze skill mastery over time using four parameters [26]. Extensions of this model include Multi-Skill Bayesian Knowledge Tracing (MS-BKT) and Contextual Guess and Slip Bayesian Knowledge Tracing (CGS-BKT). These extensions simultaneously model multiple skills and integrate contextual factors [27], [28]. Similarly, Dynamic Bayesian Networks (DBNs) could capture temporal dependencies [29].

During this period, hybrid models combining LR and psychometrics emerged. Models such as Learning Factor Analysis (LFA) use LR to examine skill mastery over time. The Additive Factor Model (AFM) uses multiple factors within a logistic framework [30]. Furthermore, Performance Factor Analysis (PFA) uses performance factors to improve prediction [31].

Traditional machine learning models, such as Support Vector Machines (SVMs), have been used for classification tasks [32]. Algorithms such as Random Forests (RF) improve overall modeling prediction [33]. Recommender systems, such as Collaborative Filtering (CF) and Probabilistic Matrix Factorization (PMF), have helped analyze learner-item interactions [34], [35].

Regarding the early evolution of reinforcement learning, Greedy and Random algorithms were established for content selection through temporal decision-making within KT [36], [37].

Additionally, key features such as personalization and knowledge decay were explored by models like Individualized Bayesian Knowledge Tracing (IBKT) [38] and BKT with Forgetting [39].

Overall, many advancements took place between 1990 and 2014. These included the core Bayesian model and its variants, hybrid models of LR and psychometrics, traditional machine learning models, and early reinforcement learning techniques. These models dynamically tracked skills by modeling multiple skills and contextual factors as temporal dependencies, thereby making predictions more accurate and personalized. This period laid a strong foundation for the evolution of deep learning models.

3) DEEP LEARNING SURGE (2015–2019)

From 2015 to 2019, the primary focus of researchers aiming to capture complex sequential patterns was on deep learning architectures. Recurrent neural networks (RNNs) formed the basis of this development. A notable model is DKT, which employs a Long Short-Term Memory (LSTM) architecture to enable temporal/sequential learning [40]. Another variant of DKT, Enhanced Deep Knowledge Tracing (DKT+), was analyzed by incorporating feature sets [41]. Another variant, DKT-Forget, incorporates knowledge decay over time [42]. Meanwhile, early deep learning models such as Neural Matrix Factorization (NMF) successfully reinforced the concept of factorization for learning analytics [43].

Similarly, exercise-aware models, such as the Exercise-Enhanced Recurrent Neural Network (EERNN) [44] and the Exercise-Aware Knowledge Tracing (EKT) [45], successfully incorporated item-related features into sequential/temporal modeling, thereby improving predictions.

Memory-augmented networks played a key role. Dynamic Key-Value Memory Networks (DKVMN), a model-like network, captured the learner's states via external memory structures [46]. Another type, Sequential Key-Value Memory Networks (SKVMN), made further advancements in capability [47].

Reinforcement learning techniques, such as the policy-based Advantage Actor-Critic (A2C) method, were used to evaluate learners' KT [48]. Meanwhile, Proximal Policy Optimization (PPO) [49] advanced to more sophisticated hybrid models, such as the Knowledge Tracing–based Knowledge Demand Model (KT–KDM). This model combines deep sequential modeling and reinforcement learning [50].

Graph-based methodologies outperformed other models. The initial model, Graph-based Knowledge Tracing (GKT), used graph-based neural networks to project inter-skill relationships [51]. Lastly, Deep Item Response Theory (DeepIRT) combines deep learning and a traditional psychometric model to integrate interpretability and prediction accuracy [52].

From 2015 to 2019, significant changes occurred in the KT world with the introduction of models capable of analyzing complex patterns and non-linear relationships. RNN variants, memory-augmented networks, reinforcement learning techniques, and graph-based models demonstrated accuracy in prediction, interpretability, and personalization. Some models could represent dynamic learner states and knowledge decay during evaluation. This made them more adaptive and advanced their ability to predict learning outcomes, eventually paving the way for more innovative KT models.

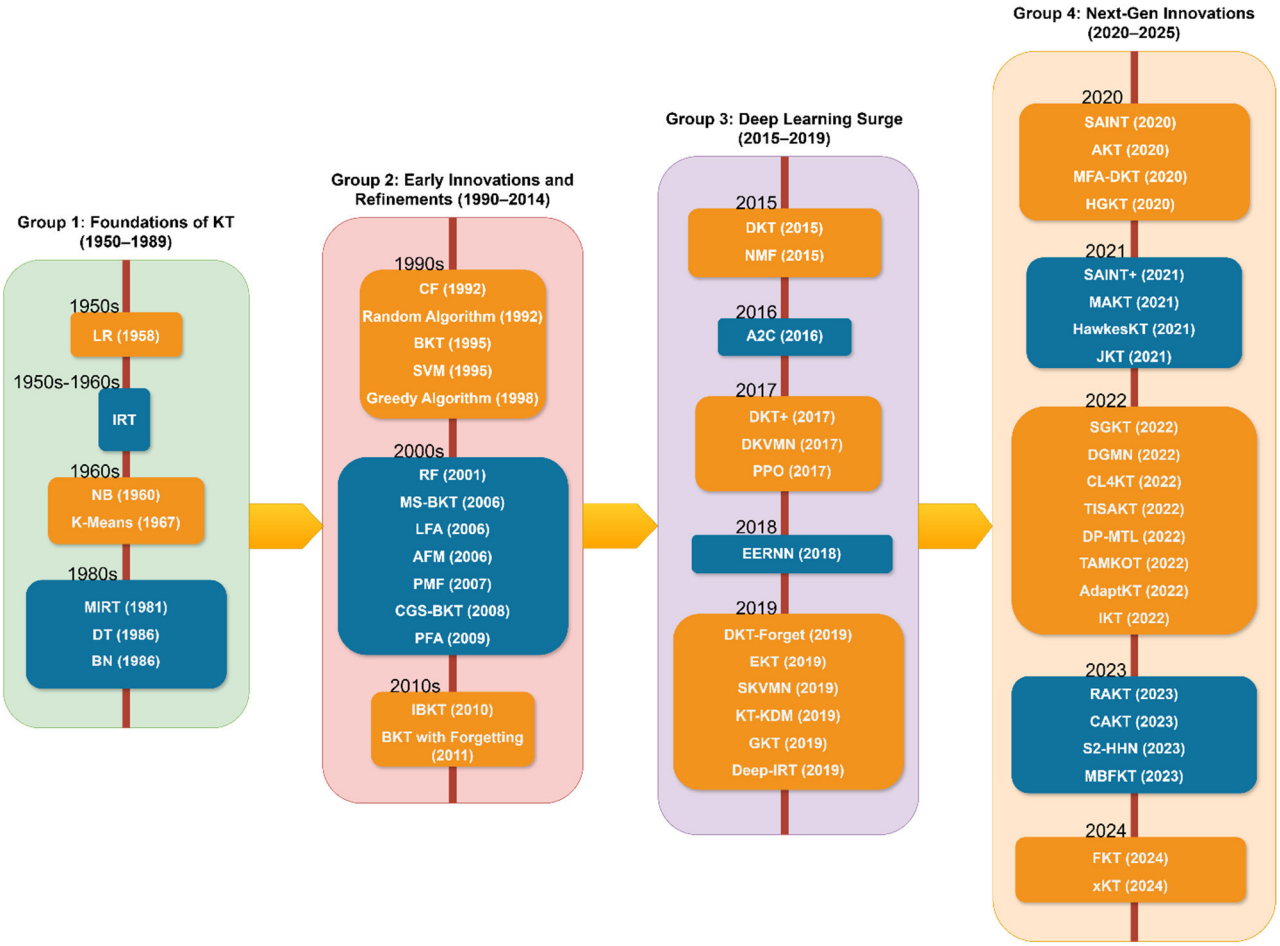

FIGURE 3. Historical overview of KT models.

4) NEXT-GEN INNOVATIONS OF KT (2020–2025)

From 2020 to 2025, the next generation of KT models advanced in several areas: attention mechanisms, Graph Neural Networks (GNNs), temporal and contrastive extensions, self-supervised learning, multi-task models, domain adaptation, and explainability.

Transformer-based architectures, such as Separated Self-Attentive Neural Knowledge Tracing (SAINT), use transformers to capture complex contextual relationships [53]. Its variant, SAINT+, incorporates a temporal feature to enhance prediction [54]. Attention mechanism models, such as Attentive Knowledge Tracing (AKT), use self-attention to analyze learner interactions [55], while Memory-Aware Knowledge Tracing (MAKT) utilizes forgetting and memory signals [56]. Relational Attention Knowledge Tracing (RAKT) focuses on inter-skill/item correlation with the help of the relational attention process [57]. Multiple Features Fusion Attention-Enhanced Deep Knowledge Tracing (MFA-DKT) uses a variety of student and item features in an attention-based procedure with DKT as its core foundation [58].

The graph-based approach includes Joint Graph Convolutional Knowledge Tracing (JKT), which uses a GNN to model skill-item relationships [59], and Hierarchical Graph Knowledge Tracing (HGKT), which shows exercise-skill hierarchies at multiple levels [60]. Session Graph Knowledge Tracing (SGKT) presents contextual variations in the form of short-term, session-based graphs [61]. Lastly, the Deep Graph Memory Network (DGMN) uses a GNN architecture with a memory feature to model forgetting in knowledge tracing [62].

Among the temporal and sequential models, Time Interval Aware Self-Attention Knowledge Tracing (TISAKT) models time gaps within the learning process [63], and Hawkes Process Knowledge Tracing (HawkesKT), a temporal point process–based model, represents the student's event-driven interactions [64]. In the contrastive learning model, Contrastive Attention Knowledge Tracing (CAKT) employs attention networks and contrastive learning to represent latent traits [65]. Similarly, the Contrastive Learning Framework for Knowledge Tracing (CL4KT) uses contrastive signals to improve KT generalization [66]. Some self-supervised

TABLE 2. Performance of the Rasch 1PL model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.4223	0.3986	0.5386	0.4957	7.3212	1
DigiArvi 2025	0.3734	0.3310	0.6147	0.4929	8.6573	1

models, such as the Self-Supervised Heterogeneous Hypergraph Network (S2-HHN), enhance self-supervision by processing heterogeneous data points that include both students and their skills [67].

Multi-task models, such as Dichotomous-Polytomous Multi-Task Learning (DP-MTL), use a framework for option tracing to provide a comprehensive overview of mastery [8]. In relation to auxiliary tasks, FKT uses response speed to capture more accurate performance metrics [68]. Meanwhile, Multiple Behavioral Features for Knowledge Tracing (MBFKT), a combination of multi-head attention, memory networks, and RNNs, precisely models multiple behavioral features [69]. The multi-activity-based model, Transition-Aware Multi-Activity Knowledge Tracing (TAMKOT), learns from transitions across activities [70]. Furthermore, Adaptable Knowledge Tracing (AdaptKT) enhances KT models to adapt across different datasets and contexts under domain adaptability [9].

Lastly, Explainable Knowledge Tracing (xKT) uses cognitive learning theories to make predictions about multiple concept questions more transparent [71]. Similarly, Interpretable Knowledge Tracing (IKT) utilizes more causal and streamlined modeled architectures for clear learner analysis [72].

This next-generation approach is significant due to its ability to model complex relationships and temporal dynamics and to perform multiple tasks simultaneously across various educational domains. The integration of GNN architectures, advanced attention mechanisms, and contrastive/self-supervised learning enhances accuracy, interpretability, and domain adaptability. Ultimately, these methods bridge the gap between theoretical constructs and practical application, paving the way for more robust and autonomous KT models in the future.

Fig. 3 provides a visual summary of the architectural evolution across these four distinct eras. In brief, from 1950 to 1989, basic statistical and probabilistic algorithms evolved into psychometric models. From 1990 to 2014, dynamic Bayesian models and hybrid logistic–psychometric methods emerged for tracking students dynamically. From 2015 to 2019, deep learning surged and introduced RNNs, memory networks, and reinforcement learning. Finally, from 2020 to 2025, sophisticated attention mechanisms, graph neural networks, and self-supervised learning were introduced, making modern predictions highly accurate, adaptive, and personalized.

C. CATEGORIES OF KT MODELS (RQ3)

This section provides a detailed discussion of various KT model categories, accompanied by a selection of representative models from each category. These models are evaluated based on their performance across two datasets: ASSISTments 09-10 and DigiArvi 2025. The models are classified into nine broad categories based on the current trends and relevance, which are outlined below:

1) PSYCHOMETRIC MODELS

Psychometric models are essential for understanding and evaluating the performance of test items and the abilities of learners. To improve understanding, these models have been broadly categorized into two subtypes: classical psychometric models (IRT and MIRT) and hybrid logistic-psychometric models (PFA, LFA, and AFM).

Classical psychometric models primarily focus on measuring latent traits and item responses. One example is IRT, which is helpful for dichotomous items (e.g., true/false) and can also be applied to polytomous items (e.g., multiple-choice) using logistic or probit functions [11]. It is important to acknowledge that classical IRT provides an estimate of a learner's ability at a single point in time. Therefore, it does not inherently capture the dynamic, evolving nature of student knowledge that is central to traditional knowledge tracing. Nevertheless, IRT is included in this taxonomy because it provides the fundamental mathematical basis on which modern tracking systems are built. Without the core parameters of item difficulty and static latent ability established by IRT, subsequent temporal extensions, such as TIRT, would not exist. MIRT is an extension of IRT that simultaneously evaluates multiple latent traits, making it useful for complex assessments that evaluate multiple skills [73].

The general formula for psychometric models is given by

$$P\left(X_{ij}=1 \mid \theta_j, a_i, b_i, c_i\right) = c_i + (1-c_i)\,\frac{1}{1+e^{-(a_i\theta_j - b_i)}} \tag{1}$$

where X_{ij} is the response of learner j on item i (correct=1 and incorrect=0), θ_j is the learner's latent ability (vector for MIRT and a single dimension for IRT), a_i is the item discriminator parameter (vector for MIRT and scalar for IRT), b_i is item difficulty, c_i is a guessing parameter. For 2PL model $c_i = 0$ and for 1PL both $c_i = 0$ and $a_i = 1$.

Hybrid logistic-psychometric models combine LR with psychometric models, such as LFA, which operate on the

TABLE 3. Performance of the BKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6853	0.6886	0.7262	0.7762	0.6210	6
DigiArvi 2025	0.7588	0.6947	0.7935	0.8319	0.5727	3

same principle. These models identify factors that influence learning and predict the future based on past data. PFA is similar to LFA and is an extension of it. PFA focuses more on the performance side of the learning process [74]. An AFM considers learning as an additive process that increments the learner's knowledge or skill level with each opportunity [75].

The formula for a hybrid logistic-psychometric model is given by

$$\left(X_{ijt} = 1 \mid \propto_i, \beta_i, F_{jt}, \rho_j, \delta_i\right) = \frac{1}{1 + e^{-(\propto_i + \beta_i F_{jt} + \rho_j + \delta_i)}} \quad (2)$$

where X_{ijt} is the response of learner j on item i at the attempt t (correct=1 and incorrect=0), $\propto_i$ difficulty level of item i, β_i is a vector coefficient capturing learning progression for linear, logarithmic, success/failure. F_{jt} is a vector for historical features that are learner-specific (success counts, failure counts, attempt numbers, and log attempts), ρ_j is a learner-specific ability parameter (LFA only), and δ_i item-specific adjustment parameter (LFA only).

For the empirical evaluation, we selected the simplest IRT formulation, the Rasch 1PL model, to predict the probability that a student will answer the next item correctly. The Rasch model assumes a single latent ability parameter for each learner and a single difficulty parameter for each item.

Table 2 shows the performance of the Rasch 1PL model across two datasets: ASSISTments 09–10 and DigiArvi 2025. Both datasets achieved notably low accuracy values compared to the standard benchmark of 0.7: 0.4223 for ASSISTments and 0.3734 for DigiArvi. This suggests that the model cannot reliably predict students' responses. The ROC-AUC scores fall significantly below 0.7 (0.3986 for ASSISTments and 0.3310 for DigiArvi), highlighting the model's difficulty in distinguishing between correct and incorrect answers and its subsequent poor performance in classification tasks. This worse-than-random performance occurs because applying a purely static, single-parameter model to highly sequential, dynamic learning data causes severe underfitting and structural model mismatch, which can lead to inverted probability assignments. Similarly, the AP values fall below the desired range of 0.7 to 0.8 (0.5386 and 0.6147, respectively), suggesting that the model underperforms in predicting correct answers. The F1 scores (0.4957 for ASSISTments and 0.4929 for DigiArvi) reflect an imbalanced precision–recall ratio, suggesting frequent false positives and false negatives. Furthermore, the high log loss values (7.3212 for ASSISTments and 8.6573 for DigiArvi) suggest overconfidence in incorrect predictions and indicate poor calibration. Finally, an overall rank score of 1 for both datasets indicates that the Rasch 1PL model performs the worst across all metrics (where 1 is the lowest rank score and 9 is the highest), as detailed in Section III-D3 and Fig. 4. This low ranking is likely due to the model's structural simplicity. However, the Rasch 1PL model's primary strength lies in its interpretability. Educators can easily understand the learned ability and item difficulty parameters, which provide valuable diagnostic insights even when predictive accuracy is poor.

2) BAYESIAN MODELS

Bayesian models are widely used in data mining related to learning analytics, particularly in KT, where learners' knowledge is evaluated over time. These models fall into one of four categories: standard (BKT), enhanced (CGS-BKT, IBKT, and MS-BKT), dynamic (Dynamic Bayesian Knowledge Tracing (DBKT) and DBN), and advanced (Hierarchical Knowledge Tracing (HKT)).

Standard Bayesian models, such as BKT (a probabilistic model), contain a basic framework in which they evaluate the learner's knowledge level as a hidden binary variable. These levels are updated based on observed responses. These models also use transition learning as a parameter for probability between levels, which is often modeled as a hidden Markov model. However, the BKT considers latent knowledge states to be static and impervious to forgetting. This can eventually lead to inaccurate parameter estimation and high computational costs [10].

In the enhanced Bayesian model, the foundation model is improved to account for contextual factors, individual differences, and multiple skills. CGS-BKT builds contextual factors into guess and slip parameters, providing a more accurate prediction of student performance by evaluating the context of each question. IBKT classifies students dynamically based on personalized question-answer data, improving prediction accuracy. MS-BKT uses multiple skills simultaneously to assess students' levels across different domains [76].

Similarly, the dynamic Bayesian model accounts for the evolution of knowledge over time, making it more temporal. Models such as DBKT use the same framework to evaluate learners' knowledge levels over time. DBNs handle dependencies between multiple variables over time, providing a more detailed view of the learning process [1], [2].

Advanced Bayesian models incorporate the multivariate Hawkes process to capture the self-exciting nature of states.

TABLE 4. Performance of the LR model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6955	0.7446	0.7967	0.7535	0.5746	7
DigiArvi 2025	0.7620	0.7981	0.8847	0.8246	0.5411	7

This type of modeling helps to further map future knowledge states based on past learning events, making KT more dynamic and interconnected [77].

The generalized formula for the Bayesian model can be divided into three parts

a) Observation model:

$$\pi\left(O \mid L_n, \theta, \phi\right) = \begin{cases} L_n \cdot \left[1 - f_S\left(\theta, \phi\right)\right] + \left(1 - L_n\right) \cdot f_G\left(\theta, \phi\right), \\ \qquad if\ O = Correct \\ L_n \cdot f_S\left(\theta, \phi\right) + \left(1 - L_n\right) \cdot \left[1 - f_G\left(\theta, \phi\right)\right], \\ \qquad if\ O = Incorrect \end{cases} \tag{3}$$

b) Bayesian Update:

$$P\left(L_n \mid O, \theta, \phi\right) = \frac{P\left(L_n\right) \cdot \pi\left(O \mid L_n, \theta, \phi\right)}{P\left(L_n\right) \cdot \pi\left(O \mid L_n, \theta, \phi\right) + \left[1 - P\left(L_n\right)\right] \cdot \pi\left(O \mid -L_n, \theta, \phi\right)} \tag{4}$$

c) Transition Learning Update:

$$P\left(L_{n+1}\right) = P\left(L_n \mid O, \theta, \phi\right) + \left[1 - P\left(L_n \mid O, \theta, \phi\right)\right] \cdot \psi\left(T, t, \phi\right) \tag{5}$$

where L_n is the latent knowledge state at n^{th} step, which ideally means the probability that the student knows the concept in a certain step. O is the response, whether it is correct or incorrect. θ is a set of contextual variables such as item difficulty or presentation format, ϕ are additional parameters such as differences within individuals, skill-specific traits, and hierarchical factors, $f_S\left(\theta, \phi\right)$ is a slip function that ideally states the probability of making an error while knowing the concept. $f_G\left(\theta, \phi\right)$ is a guess function, which means the probability of guessing correctly without knowing the concept, $P\left(L_n\right)$ is the probability of knowing a concept before a certain observation. $\pi\left(O \mid L_n, \theta, \phi\right)$ is the probability of response O in a latent knowledge state L_n, $\pi\left(O \mid -L_n, \theta, \phi\right)$ is the probability of response O when the student does not know the concept area, T is the learning rate, t is the time gap between learning events, and $\psi\left(T, t, \phi\right)$ is a transition learning function.

For the empirical evaluation, we implemented the standard BKT model. Unlike more advanced extensions, the standard version assumes independence between skills, does not account for forgetting, and uses only slip and guess parameters to model error and noise.

Table 3 shows that BKT has different levels of predictive performance. On ASSISTments, BKT achieves an overall rank score of 6, outperforming simpler baselines such as Rasch 1PL and some advanced models. However, on DigiArvi, BKT drops to an overall rank score of 3 (third lowest overall), falling behind advanced models such as DKT, GKT, and FKT (refer to Section III-D3 for the detailed ranking). Despite this inconsistency, BKT's strength lies in its interpretability. Its slip, guess, and transition parameters offer clear pedagogical insights, enabling educators to design targeted interventions even when its predictive performance is weaker than that of newer models.

3) MACHINE LEARNING MODELS

Machine learning models can learn from labeled data and provide detailed insights. Some can discover innovative patterns in unlabeled data, while others can learn from prior feedback to improve actions, which is essential for KT. These models are primarily divided into three categories under KT: supervised learning (DT, RF, SVM, LR, NB, and BN); unsupervised learning (K-means, PMF, and CF); and reinforcement learning (PPO, A2C, Greedy, and KT-KDM).

Supervised learning models are trained on labeled datasets to predict outcomes or perform classification. A DT splits data into branches to make predictions. DT models are useful for classification tasks but can be prone to overfitting when analyzing complex datasets [4]. RF, on the other hand, is an extension of DT. It is an ensemble method in which analysis is performed based on multiple DTs. The results are then merged for enhanced accuracy compared to DT. RF is also less prone to overfitting [5]. The SVM model excels at classification tasks, finding an appropriate hyperplane that best defines the classes [3]. LR and NB models stem from early statistical and probabilistic models. LR is simple and effective for binary classification tasks, while NB is related to probabilistic classification [5], [78]. BNs are graphical models that depict probabilistic relationships between parameters [79].

The following formula can describe this type of model:

$$\min_{\theta} \sum_{i=1}^{N} \left[L\left(y_i, f\left(x_i; \theta\right)\right) + \lambda R\left(\theta\right)\right] \tag{6}$$

where x_i are the input features, y_i are the known labels or the targets, θ is the parameter, $f\left(x_i; \theta\right)$ is the prediction for the model, L is a loss function within the model while a prediction is made, $R\left(\theta\right)$ is the regularization term within the model for preventing overfitting, λ is hyperparameter tuning for the model to get better predictions, N is the number of students.

TABLE 5. Performance of the DKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6636	0.6387	0.6968	0.7740	0.6252	5
DigiArvi 2025	0.7443	0.7412	0.8503	0.8279	0.5313	6

In unsupervised learning, pattern identification or grouping is based on unlabeled data. Algorithms like K-means are known for clustering and partitioning data into K clusters to improve understanding [10]. PMF and CF are components of recommender systems. PMF is primarily used for dimensionality reduction, and CF predicts users' preferences based on past data [8].

The general formula for an unsupervised model is given by

$$\min_{\theta,Z} \sum_{i=1}^{N} \left[D\left(x_i, g\left(z_i; \theta\right)\right) + \lambda R\left(\theta, Z\right) \right] \tag{7}$$

where x_i are the observed unlabeled data, z_i are the hidden representations, $g\left(z_i; \theta\right)$ is the reconstruction or, we can also say, the generative function, D is the divergence measure or simply the distance within the model, $R\left(\theta, Z\right)$ is the regularization term within the model for preventing overfitting. λ, N and θ is the same as from equation 6.

Lastly, in reinforcement learning, the agent learns through interaction with the environment. PPO (policy-based) is one of the most popular algorithms in reinforcement learning and is used extensively in gaming, robotics, and control systems in real-world scenarios [80]. A greedy algorithm selects the action based on the highest rewards, while A2C, an extension of PPO, combines policy-based and value-based approaches. Furthermore, KT-KDM is a more specific model that combines KT with reinforcement learning to generate personalized learning paths for students [50].

Reinforcement learning can be formulated as follows:

$$\max_{\pi} \mathrm{E}_{\tau\sim\pi} \left[\sum_{t=0}^{T} \gamma^t r\left(s_t, a_t\right) \right] \tag{8}$$

where π is the policy (or strategy for the action), s_t is the state at time t, a_t is the action at time t, $r\left(s_t, a_t\right)$ is the reward received after taking action a_t at state s_t, γ^t is the discount factor that decides how much future reward is valued, τ are the trajectories $(s_0, a_0, s_1, a_1, \ldots)$ for the policy π, $\mathrm{E}_{\tau\sim\pi}$ is an expectation over the trajectories.

For the empirical evaluation, we implemented LR as the representative supervised machine learning model. LR provides a strong and transparent baseline for KT tasks. Unlike more complex methods, LR does not directly model temporal dependencies, yet it can capture significant patterns when features such as student, item, and skill are included. Additionally, LR supports interpretability by assigning weights to input features.

As shown in Table 4, LR delivers strong performance consistently across datasets. On ASSISTments, LR achieves an overall rank score of 7 (3rd best overall), just behind TIRT and AdaptKT. This highlights the fact that even a simple machine learning baseline can outperform several Bayesian and graph-based approaches. On DigiArvi, LR also achieves an overall rank score of 7, slightly behind FKT and TIRT. However, it outperforms traditional baselines like Rasch 1PL in terms of calibration and stability (see Section III-D3 for detailed rankings). These results reinforce LR's effectiveness as a reliable, interpretable baseline in KT research that strikes a balance between predictive performance and model transparency.

4) DEEP LEARNING MODELS

Models of deep learning for knowledge tracing are part of the broader machine learning category. These models have drastically transformed learner analysis by capturing complex patterns. Major groups include RNN-based KT (DKT, LSTM, Bi-LSTM, NMF, EERNN, EKT, and DKT+); memory-augmented models (DKVMN and SKVMN); attention models (AKT, MAKT, RAKT, and MFA-DKT); hybrid models (Deep-IRT, xKT, and IKT); and transformer-based KT (SAINT, SAINT+, and Student State-Aware Knowledge Tracing (SSKT)).

RNN-based DKT models predict students' future performance based on past data through RNNs. DKT is one of the earliest RNN-based models to use LSTM networks to capture learner traits and predict future performance based on past responses. LSTM is a network model that forms the basis of many KT models and captures long-term dependencies. Its extension, Bi-LSTM, captures both forward and backward dependencies, making it more accurate [81]. When a neural factor is added to matrix factorization, NMF enhances the ability to capture patterns that traditional MF struggles with [6]. EERNN and EKT embed exercise-related information into the RNN framework to differentiate between exercises and their associated KCs. They also use exercise difficulty and similarity concepts to improve prediction accuracy [82]. DKT+ is an enhanced version of DKT that incorporates attention layers or personalized learning rates to improve model interpretability [7].

To understand this process intuitively, imagine an RNN acting as a teacher observing a student taking an ongoing exam. As the student answers each question, the teacher updates their mental summary of the student's overall understanding (this represents the network's 'hidden state'). An LSTM improves upon this by learning what information to 'remember' and what to 'forget' over time. For example, it recognizes when a student consistently grasps a core

concept and when they make a one-off careless mistake. Memory-augmented models take this analogy a step further; they act like a teacher with a highly organized physical grade book (the external memory matrix). Rather than relying on a single mental summary, the model can independently update the student's level of mastery for specific topics (e.g., algebra versus geometry).

The formula for this group of models is given by

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L\left(y_{i,t}, \sigma\left(W \cdot RNN\left(x_{i,t}\right) + b\right)\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (9)$$

where T_i is the number of time steps for student i, $x_{i,t}$ is the input parameter for student i at time t, $y_{i,t}$ are the known labels or targets at time t, W is the weight of the output layer, and b is its bias. σ is the nonlinear activation function, $RNN\left(x_{i,t}\right)$ is the hidden state representation of knowledge at time t, $\sigma(W \cdot RNN(x_{i,t}) + b$ is the probability that the student answers correctly at time t. $\theta, N, R\left(\theta\right), L$ and λ are the same as equation 6.

Memory-augmented models have memory mechanisms in place to improve efficiency and accuracy. Models like DKVMN use key-value pairs to represent knowledge states. The key in the matrix is static and stores the representation of KCs, while the value in the matrix is dynamic and shows the mastery level of each KC. SKVMN addresses DKVMN's limitations by using a modified LSTM for sequential modeling of the key-value memory network [1].

The formula for memory-augmented models is as follows

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L\left(\begin{matrix} y_{i,t}, MemoryRead \\ \left(MemoryUpdate\left(M_{i,t}, x_{i,t}\right)\right) \end{matrix}\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (10)$$

where $M_{i,t}$ is the memory matrix representing student i at time t, *MemoryUpdate* is for the operation to write new information, *MemoryRead* is the operation for extracting the current input from the knowledge. $\theta, N, R\left(\theta\right), L$ and λ are the same as equation 6. $T_i, x_{i,t}$ and $y_{i,t}$ are the same as equation 9.

Attention-based models improve learner performance by considering temporal, relational, and multi-feature insights when modeling knowledge. AKT has a monotonic attention mechanism that reduces the attention weight for questions based on their time distance. Meanwhile, MAKT uses self-attention (short-term memory) and a key-value memory network (long-term memory) to update the learner's knowledge states, similar to DKVMN. This makes MAKT more appropriate for evaluation. RAKT combines attention weights (like AKT) with relation coefficients derived from exercise relation modeling. This model essentially uses text information from the questions to interpret their relationship [1]. MFA-DKT concatenates multiple student behavior and exercise features into the attention mechanism so it can assign different weights to the features, resulting in superior learning prediction performance [58].

The formula for attention-based models is as follows:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L\left(y_{i,t}, \sigma\left(W \cdot Attention\left(x_{i,t}\right) + b\right)\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (11)$$

where *Attention* is the attention mechanism to weigh past data. $\theta, N, R\left(\theta\right), L$ and λ are the same as equation 6. $T_i, x_{i,t}, y_{i,t}, W, b$ and σ are the same as equation 9.

Hybrid deep learning models, such as Deep-IRT, combine deep learning features with psychometric principles. This improves the predictability of learner data, making it more interpretable and accurate [73].

The formulation of deep-IRT can be stated as follows:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \begin{bmatrix} L\left(\begin{matrix} y_{i,j}, c_j + \left(1 - c_j\right) \\ \cdot\sigma\left(-\left(a_j\theta_i - b_j\right)\right) \end{matrix}\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (12)$$

where Q_i is the number of questions attempted by student i, $y_{i,j}$ is the true response, a_j is the item discriminator parameter, b_j is item difficulty, c_j is a guessing parameter. $\theta, N, R\left(\theta\right), L$ and λ are the same as equation 6. W, b and σ are the same as equation 9.

Explainable and interpretable hybrid models, which are widely used in deep learning models but can also be used in different KT model categories, help educators understand the learning process. xKT models aim to provide a clear explanation of how predictions are made, while IKT ensures that the reasons behind certain predictions are understood [74].

$$\min_{\theta} \sum_{i=1}^{N} \begin{bmatrix} L\left(y_i, f\left(x_i; \theta\right)\right) + \gamma \cdot L_{expl}\left(E_i\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (13)$$

where γ is the weighting coefficient for explanation loss, E_i is the explanation output at time t, L_{expl} is the explanation loss. $\theta, N, R\left(\theta\right), L, \lambda, x_i, y_i$ and $f\left(x_i; \theta\right)$ are the same as equation 6.

Transformer-based models related to deep learning are highly advanced, enabling them to capture the complex dependencies of learners. SAINT is a transformer-based model that uses an encoder-decoder framework for predictions. The encoder processes the sequence of exercises, and the decoder handles the sequence of responses. SAINT+ improves upon the attention mechanism for interpretability and has a temporal feature that captures learning traits [68]. SSKT also uses an attention mechanism to integrate information related to new exercises and combines an LSTM for sequence coding [83].

T he formula for this transformer-based model is as follows:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L\left(y_{i,t}, \sigma\left(W \cdot h_{i,t} + b\right)\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (14)$$

$$\mathrm{h_{i,t}} = \begin{cases} TransformerDecoder\left(x_{i,t}, Encoder\left(q_{i,t}\right)\right), & SAINT \\ Attention\left(q_{i,t}, r_{i,t}\right), & SAINT+ \\ Concat\left(LSTM\left(x_{i,t}\right), Transformer\left(x_{i,t}\right)\right), & SSKT \end{cases} \quad (15)$$

TABLE 6. Performance of the GKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6255	0.5670	0.6369	0.7622	0.6639	3
DigiArvi 2025	0.7313	0.7118	0.8252	0.8332	0.5669	4

where $h_{i,t}$ is the hidden state representation of knowledge at time t, $r_{i,t}$ is the sequence of responses, $q_{i,t}$ is the sequence of questions attempted. θ, N, $R(\theta)$, L and λ are the same as equation 6. T_i, $x_{i,t}$, $y_{i,t}$, W, b and σ are the same as equation 9.

For the empirical evaluation, we implemented DKT as the representative deep learning-based KT model. DKT models use deep neural networks to capture complex, nonlinear patterns in student learning data. While these models are highly effective in predictive tasks, they are less interpretable than traditional models.

Table 5 shows that deep learning models perform moderately well across datasets. On the ASSISTments 09–10 dataset, for example, they achieve an overall rank score of 5, with an F1 score of 0.7740 and an ROC-AUC of 0.6387, demonstrating moderate predictive ability. On DigiArvi 2025, they achieved an overall rank score of 6, with an accuracy of 0.7443, an F1 score of 0.8279, and an AP of 0.8503. These results confirm the effectiveness of deep learning–based KT models, though they remain less interpretable than simpler KT families.

5) GRAPH MODELS

Graph-based models are part of the broader machine learning category. These models are innovative because they use graph structures to predict the learning process. The three main subclasses within this group are core graph models (GKT), GNN-based models (HGKT and JKT), and advanced graph models (SGKT and DGMN).

Core graph models have laid the foundation for more advanced models to build upon. In GKT, the nodes are the KCs, the edges represent the dependency relation between the KCs, and the graphs are constructed based on sequences of student interactions [51].

The formula for the GKT model is as follows:

$$\min_{\theta} \sum_{i=1}^{N} \left[L\left(y_i, \sigma\left(\sum_{u \in N(v)} W \cdot h_{u,i} + b\right)\right) + \lambda R(\theta) \right] \quad (16)$$

where $h_{u,i}$ is the hidden state representation of knowledge of the neighbor node u for learner i, $N(v)$ is the set of neighbor nodes of the target node v. θ, N, $R(\theta)$, L, λ and y_i are the same as equation 6. W, b and σ are the same as equation 9.

GNN-based models, such as HGKT, build upon the core model and capture the hierarchical evolution of knowledge using graph structures [60]. These models also represent the relationship between different KCs, making them more predictable. Conversely, JKT uses Graph Convolutional Networks (GCNs) to apprehend higher-order information within the knowledge graph. This allows for the aggregation of data from neighboring nodes, providing a better understanding of knowledge states and temporal effects [59].

The generalized formulation for the GNN-based model is given by

$$\min_{\theta} \sum_{i=1}^{N} \left[L\left(y_i, \sigma\left(\sum_{u \in N(v)} \propto_{vu} \cdot W \cdot h_{u,i} + b\right)\right) + \lambda R(\theta) \right] \quad (17)$$

where $\propto_{vu}$ is the attention or weighing coefficient between nodes v and u, capturing relations such as hierarchical/convolutional structure. θ, N, $R(\theta)$, L, λ and y_i are the same as equation 6. W, b and σ are the same as equation 9. $h_{u,i}$ and $N(v)$ are the same as equation 16.

Advanced graph models can integrate memory networks and session-based models, making them more effective for prediction. SGKT uses session-based learning in graph-based models, allowing for the dynamic adaptation of knowledge graphs based on session-specific interactions. This approach also helps capture the temporal dynamics of sessions [61]. Similarly, DGMN integrates memory networks into the graph structure and focuses on the forgetting mechanism. This model captures the knowledge state of memory across KCs and uses an attention mechanism to link questions with their KCs [1].

The generalized formulation for the advanced graph model is given by

$$\min_{\theta} \sum_{i=1}^{N} \left[L\left(y_i, \sigma\left(\sum_{u \in N(v)} \propto_{vu} \cdot W \cdot h_{u,i} + \gamma \cdot M\left(h_{v,i}, m_i\right) + \beta \cdot S\left(h_{v,i}, s_i\right) + b\right)\right) + \lambda R(\theta) \right] \quad (18)$$

where $h_{v,i}$ is the hidden state representation of knowledge of node v for learner i, γ is the scaling parameter for memory influence, m_i is the memory embedding for interaction i, M is the memory integration function, β is the scaling parameter for session-based influence, s_i is the session embedding, and S is the session-based integration function. θ, N, $R(\theta)$, L, λ and y_i are the same as equation 6. W, b and σ are the same as equation 9. $h_{u,i}$ and $N(v)$ are the same as equation 16. $\propto_{vu}$ is the same as equation 17.

For the empirical evaluation, we implemented GKT as the representative graph-based model. GKT uses graph structures to model relationships between skills and provide insight

TABLE 7. Performance of the TIRT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6960	0.7468	0.7997	0.7536	0.5732	9
DigiArvi 2025	0.7621	0.7991	0.8855	0.8248	0.5414	8

TABLE 8. Performance of the FKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6024	0.5676	0.6433	0.7519	0.6694	2
DigiArvi 2025	0.7708	0.7855	0.8718	0.8463	0.4988	9

into skill dependencies. Although less transparent than LR, GKT's graph structure offers interpretability by highlighting inter-skill relationships.

Table 6 shows that GKT achieves an overall rank score of 3 on the ASSISTments 09–10 dataset, with an accuracy of 0.6255 and an ROC-AUC of 0.5670. This indicates relatively weaker predictive performance compared to other KT model families. In the DigiArvi 2025 dataset, GKT achieves an overall rank score of 4, reflecting performance at the lower end of the moderate range across various metrics. Refer to Section III-D3 for detailed rankings. GKT has an accuracy of 0.7313, an F1 score of 0.8332, and an AP of 0.8252.

6) TEMPORAL AND SEQUENTIAL MODELS

Temporal and sequential models are pivotal because they help us understand how knowledge states evolve. These models can be divided into two categories: statistical/probabilistic models, such as BKT with forgetting and TIRT, and deep learning models, such as HawkesKT, DKT-Forget, and TISAKT.

Among statistical/probabilistic models, BKT with a forgetting mechanism is employed, taking into account factors such as the time since a student's last interaction and past attempts. These attributes enable the accurate estimation of student performance. TIRT is a temporal extension of IRT; therefore, it can model changes in students' abilities over time [84].

The BKT formula with forgetting can be expressed as follows:

$$P\left(K_t|x_t\right) = (1-\delta)\cdot P\left(K_{t-1}\,|\,x_{t-1}\right) + \delta\cdot\left(1-P(K_{t-1}|x_{t-1})\right) \tag{19}$$

where x_t is the student's response at time t, $P\left(K_t\right)$ is the probability that the student knows the skill at time t, $P(K_{t-1})$ is the probability that the student knows the skill at time *t-1*, δ is the forgetting rate or decay factor.

The formula for TIRT is as follows:

$$P\left(X_{ij}=1\,|\,\theta_j, a_i, b_i, c_i, \gamma_j, \Delta t\right) = c_i + (1-c_i)\,\frac{1}{1+e^{-(a_i\theta_j - b_i)}} + \gamma_j\cdot\Delta t \tag{20}$$

where γ_j is the temporal effect parameter for item *j*, capturing the influence of time in learning. Δt is the time difference between the current and previous interactions. X_{ij}, θ_j, a_i, b_i and c_i are the same as equation 1.

Within deep learning models, the Hawkes-KT model describes the temporal evolution of learners. This model assumes that mastery of a KC is influenced not only by previous interactions with that KC but also by interactions with other KCs. This makes it a model for a deeper understanding of learning dynamics [64]. DKT-Forget extends the core DKT model by incorporating features of forgetting behavior, such as time elapsed and the most recent student interaction, thereby making it more predictable. Conversely, TISAKT uses an attention mechanism to capture the influence of time intervals between learning, a factor often overlooked in traditional models [63].

The formula for deep learning based temporal models is given by

$$\min_{\theta} \sum_{i=1}^{N}\sum_{t=1}^{T_i}\left[\begin{array}{l} L\left(y_{i,t}, \sigma\left(W\cdot h_{i,t}+b\right)\right) \\ +\lambda R\left(\theta\right)\end{array}\right] \tag{21}$$

$$\mathrm{h_{i,t}} = \begin{cases} HawkesProcess\left(x_{i,t}\right), & HawkesKT \\ LSTM\left(x_{i,t}\right)\cdot Forget(x_{i,t}), & DKT-Forget \\ Attention\left(x_{i,t}, \delta_t\right), & TISAKT \end{cases} \tag{22}$$

where δ_t is the time embedding parameter. θ, N, $R\left(\theta\right)$, L and λ are the same as equation 6. T_i, $x_{i,t}$, $y_{i,t}$, W, b and σ are the same as equation 9. $h_{i,t}$ is the same as equation 14.

For the evaluation, we implemented the TIRT model as the representative temporal/sequential model. This model augments item effects with a simple within-student temporal covariate, allowing for student ability to drift over time.

Table 7 shows TIRT's performance across datasets. On the ASSISTments 09–10 dataset, TIRT achieves the highest overall rank score of 9, with a strong F1 score (0.7536) and ROC-AUC (0.7468), reflecting its ability to capture sequential patterns effectively. On DigiArvi 2025, TIRT achieves an overall rank score of 8, with an accuracy of 0.7621 and an F1

TABLE 9. Performance of the CLKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.6582	0.6844	0.7276	0.7141	0.6278	4
DigiArvi 2025	0.7554	0.7971	0.8839	0.8186	0.5418	5

TABLE 10. Performance of the AdaptKT model on datasets.

Dataset	Accuracy	ROC-AUC	Average Precision	F1 Score	Log Loss	Overall Rank Score
ASSISTments 09–10	0.7162	0.7054	0.7711	0.7935	0.6040	8
DigiArvi 2025	0.6373	0.4708	0.7723	0.7663	0.6458	2

score of 0.8248. This confirms that incorporating temporal dynamics improves the prediction of student responses across datasets.

7) MULTI-TASK MODELS

Multi-task learning models have become increasingly significant as data has become more sophisticated, creating a need for models that can handle multiple tasks simultaneously. Examples of multitask learning KT models include DP-MTL, FKT, MBFKT, and TAMKOT, all of which are part of the deep learning model category.

Models such as DP-MTL combine KT with option tracing to improve learner assessment by predicting the correctness of both the question and the options in a multiple-choice question. This framework improves the granularity of knowledge representation, leading to better performance on tasks such as score prediction [8]. FKT predicts response speed as an auxiliary task to address consistency between predicted and observed performance. It uses an encoder-decoder predictor framework to provide valuable insights [68]. MBFKT combines a multi-head attention mechanism to identify the impact of various behaviors on students' learning states. The model also incorporates memory networks and RNNs to update and track knowledge mastery dynamically [69]. TAMKOT assesses learners' knowledge states between assessed and non-assessed learning activities, allowing unlimited transitions between learning types. TAMKOT outperforms traditional models by modeling both assessed and unassessed parts, providing valuable insights into learning transitions [70].

The following formula can describe this multi-task model:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L_1\left(y_{i,t}, f\left(x_{i,t};\theta\right)\right) \\ +\sum_{k=2}^{K} \propto_k L_k\left(auxiliarytask_k\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \tag{23}$$

where L_1 is the loss function within the model to be used, $f\left(x_{i,t};\theta\right)$ is the prediction for the model, K is the task, $\propto_k$ weight for the auxiliary task, L_k is the loss function for the auxiliary task. θ, N, $R\left(\theta\right)$ and λ are the same as equation 6. T_i, $x_{i,t}$ and $y_{i,t}$ are the same as equation 9.

For the evaluation, we implemented FKT as the representative model for multi-task learning. FKT is designed to capture task-specific dynamics while maintaining shared knowledge across tasks. This improves predictive performance and generalization.

Table 8 shows FKT's performance across datasets. On the ASSISTments 09–10 dataset, FKT achieves an overall rank score of 2, with an accuracy of 0.6024 and an ROC-AUC of 0.5676. This indicates that FKT struggles to outperform simple models. In contrast, on DigiArvi 2025, FKT achieves the highest overall rank score of 9, with an accuracy of 0.7708 and an F1 score of 0.8463, demonstrating its ability to capture task-specific patterns.

8) CONTRASTIVE AND SELF-SUPERVISED MODELS

Contrastive and self-supervised KT models are built on powerful representations of latent knowledge and interactions in situations where labeled data is scarce or unavailable. These models are further divided into contrastive (CAKT, CL4KT) and self-supervised (S2-HHN) models.

Contrastive learning models, such as CAKT, use an attention mechanism to determine the importance of the questions students answer. These models also use a contrastive loss function to compare similar and dissimilar interactions. This pulls similar interactions together and pushes dissimilar ones away from the knowledge representation space [65]. CL4KT uses contrastive learning for student responses over time. It considers each student's performance on different questions, grouping them as positive or negative samples. Then it pulls similar interactions together and pushes dissimilar ones away [66].

Among self-supervised models, S2-HHN combines contrastive learning with self-supervision signals to improve graph-based KT representation. The graph represents relationships among heterogeneous entities (e.g., students, questions, or concepts), and contrastive learning helps the

model distinguish between similar and dissimilar interactions, thereby optimizing it [67].

The following formula can describe this contrastive/self-supervised model:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L_{self-supervised}\left(y_{i,t}, f\left(x_{i,t};\theta\right)\right) \\ +L_{contrastive}\left(y_{i,t}, f\left(x_{i,t};\theta\right)\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (24)$$

where $L_{self-supervised}$ is the self-supervised loss function within the model to be used, $L_{contrastive}$ is the loss function for the contrastive task. θ, N, $R(\theta)$ and λ are the same as equation 6. T_i, $x_{i,t}$ and $y_{i,t}$ are the same as equation 9. $f\left(x_{i,t};\theta\right)$ is the same as equation 23.

For the empirical evaluation, we implemented CLKT, a simplified version of CL4KT without the sequential component, as the representative contrastive/self-supervised learning model. CLKT learns item embeddings through contrastive pretraining, utilizing co-occurring items in student interaction sequences as positive examples. Then, these embeddings are fine-tuned with logistic regression to predict the correctness of student responses, integrating the learned embeddings and additional metadata.

As shown in Table 9, CLKT achieves an overall rank score of 4 on ASSISTments 09–10 and 5 on DigiArvi 2025. See Section III-D3 for the complete rankings. On ASSISTments 09–10, the model achieves an accuracy of 0.6582, an ROC-AUC of 0.6844, an AP of 0.7276, and an F1 score of 0.7141. These results indicate that the model performs decently compared to other models. On DigiArvi 2025, the model performs slightly better with an accuracy of 0.7554, an F1 score of 0.8186, and an AP score of 0.8839. This reflects good, though not top-tier, performance.

9) DOMAIN ADAPTIVE MODELS

Domain adaptation models are used to adapt KT models to new domains or datasets. The key idea is to make KT models flexible enough to adapt to different learning environments without having to retrain the entire model from scratch. A prominent model in this category is AdaptKT.

The AdaptKT model primarily operates on three dynamics. The first is dynamic adaptation, in which the KT model updates the relevant parts to adapt to new data. The second is transfer learning, in which knowledge from the previous domain is transferred to the new one. The third is flexible training, in which there is no need to start from scratch; the new training is done on top of the previous knowledge base [9].

The following formula can describe this domain adaptive model:

$$\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \begin{bmatrix} L\left(y_{i,t}, f\left(x_{i,t};\theta\right)\right) \\ +L_{adapt}\left(y_{i,t}, f\left(x_{i,t};\theta\right)\right) \\ +\lambda R\left(\theta\right) \end{bmatrix} \quad (25)$$

where L_{adapt} is the loss function for the adaptive task, which helps the model to adjust to a new domain. $\theta, N, L, R(\theta)$ and λ are the same as equation 6. T_i, $x_{i,t}$ and $y_{i,t}$ are the same as equation 9. $f\left(x_{i,t};\theta\right)$ is the same as equation 23.

For the empirical evaluation, we implemented AdaptKT as the representative domain-adaptive model. It performs unsupervised CORAL feature alignment from a source domain to a target domain. Then, it trains a logistic classifier on the aligned source and evaluates it on the aligned target. This makes AdaptKT effective for adapting models across different domains.

As shown in Table 10, AdaptKT achieves an overall rank score of 8 on ASSISTments 09–10. It achieves an accuracy of 0.7162, an F1 score of 0.7935, and an AP of 0.7711. These results indicate that AdaptKT performs better than most KT models. However, on DigiArvi 2025, AdaptKT achieves an overall rank score of 2, with lower performance compared to other models, achieving an accuracy of 0.6373, an F1 score of 0.7663, and an AP of 0.7723.

10) OVERVIEW OF THE MODELS RELATED TO KT

The KT models can be broadly categorized into nine groups: psychometric, Bayesian, machine learning, deep learning, graph-based, temporal/sequential, multi-task, contrastive/self-supervised, and domain-adaptive. The first five categories form the core KT families, while the last four build on these by adding transfer-learning, auxiliary tasks, temporal features, or domain adaptation. The shared aim is to personalize learning by improving predictive power.

Empirical testing of one representative per category reveals clear trends. Rasch 1PL (psychometric) and GKT (graph) performed the worst, demonstrating generally weaker predictive power. DKT (deep learning) provides a solid baseline, delivering consistent mid-rank performance. BKT (Bayesian) and CLKT (contrastive/self-supervised) demonstrate moderate but inconsistent improvement. Notably, models such as AdaptKT (domain-adaptive) and FKT (multi-task) exhibit highly polarized, dataset-specific behavior—excelling on one dataset but struggling on another. Ultimately, the highest and most consistent ranks are achieved by temporal/sequential and supervised machine learning models. TIRT and LR models perform consistently better than most others across various metrics.

D. EMPIRICAL ANALYSIS OF KT MODELS (RQ4)

This section presents a detailed empirical analysis of the performance of nine representative KT models from each category based on evaluation metrics, datasets, prediction curves, and student-level predictions. The analysis is divided into the following subsections:

1) HEATMAP OF MODEL RANK SCORE BY METRIC PERFORMANCE

In this section, we provide a thorough analysis of nine exemplary KT models. These models are evaluated based on various performance metrics, such as accuracy, ROC-AUC, AP, F1 score, and log loss, across two distinct datasets: ASSISTments 09-10 and DigiArvi 2025. The results are displayed as a heatmap, facilitating easy comparison of the models' relative performance.

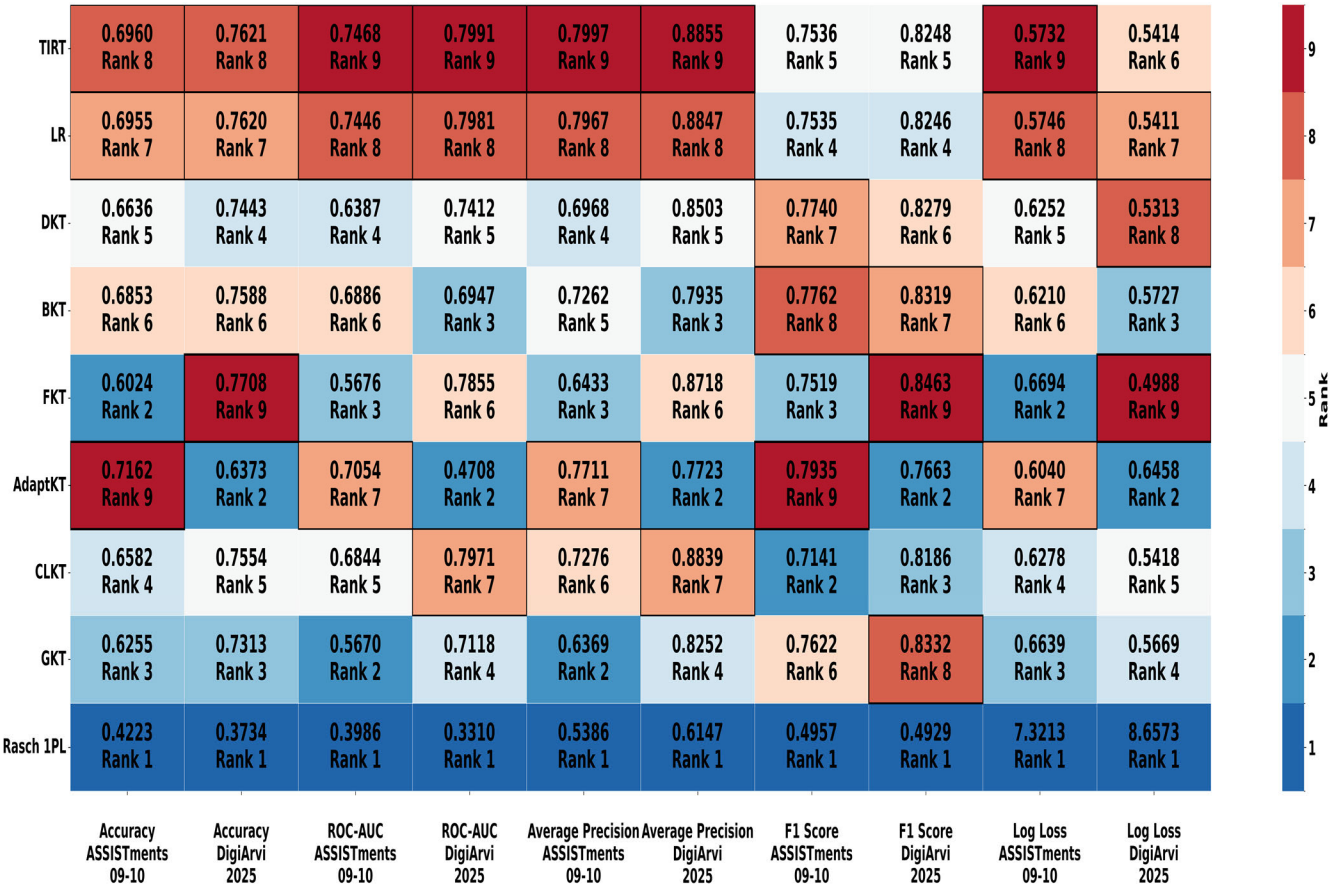

FIGURE 4. Heatmap of model rank score by metric performance.

Fig. 4 uses a color-coded scheme where shades of red indicate the highest rank score and shades of blue indicate the lowest rank score. Each cell shows a model's performance metric value and its corresponding rank score. This provides a visual summary of how each model performs across different evaluation criteria.

Upon reviewing the heatmap, we observe that TIRT performs consistently well across most of the metrics. Its high accuracy and strong performance on AP and ROC-AUC are notable, with top-tier rank scores (8 or 9) in these areas across both datasets. However, its performance is more moderate on the F1 score, achieving a rank of 5 for both datasets. Despite this dip in the F1 score, TIRT demonstrates high reliability and robust performance in predicting student proficiency overall. This model is a top contender for scenarios where high performance across various metrics is essential.

By comparison, models such as LR and DKT demonstrate notable and consistent mid-to-high performance across both datasets. Meanwhile, FKT and AdaptKT achieve top-tier rank scores on specific datasets (DigiArvi and ASSISTments, respectively) but perform significantly worse on the other dataset, lacking the cross-domain consistency of TIRT. Conversely, models such as Rasch 1PL, GKT, BKT, and CLKT tend to perform inconsistently or poorly across both datasets. Rasch 1PL achieves the lowest rank score (1) in all metrics, indicating its limited ability to capture the complexities of student learning and accurately predict future outcomes.

Based on the overall rank score and metric performance, the DigiArvi 2025 dataset generally yields better predictive performance across several models, particularly for accuracy, ROC-AUC, AP, and F1 score. While the ASSISTments 09-10 dataset performs well with models such as TIRT, the DigiArvi 2025 dataset produces stronger raw results across various key metrics, suggesting its features may be more readily captured by the models evaluated in this study.

Ultimately, the heatmap provides a clear comparison of the models' strengths and weaknesses across multiple dimensions, allowing for an informed selection of the most suitable model for a given task. It highlights the trade-offs between a model's ability to balance multiple performance metrics and confirms that TIRT is the most robust choice for applications requiring high performance across a wide range of evaluation criteria.

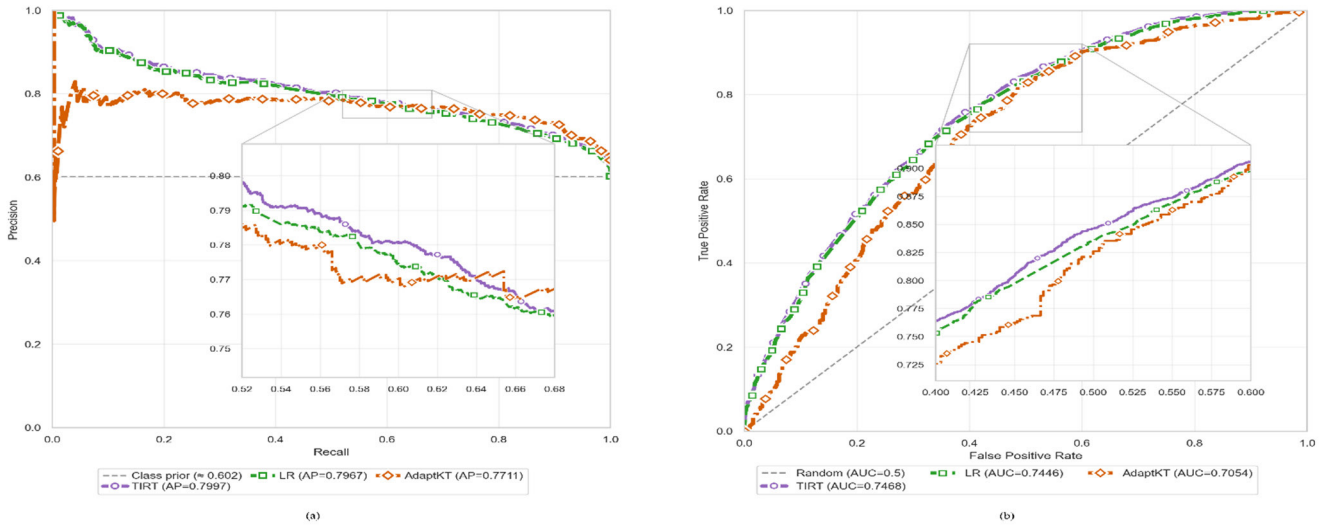

FIGURE 5. PR and ROC curves for the top 3 KT models in ASSISTments 09-10.

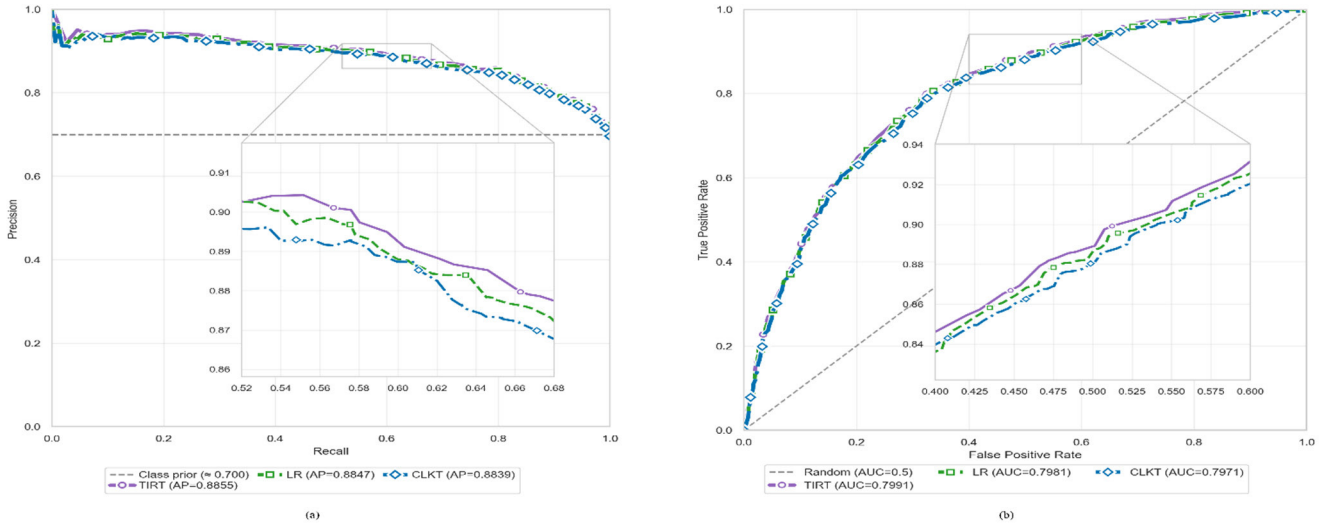

FIGURE 6. PR and ROC curves for the top 3 KT models in DigiArvi 2025.

2) CURVES FOR KEY MODELS

This section compares the performance of the top KT models across the ASSISTments 09-10 and DigiArvi 2025 datasets. It uses PR and ROC curves to evaluate how well the models balance precision, recall, and discriminatory power.

Fig. 5(a) shows the PR curves for the top three KT models on the ASSISTments dataset: TIRT, LR, AdaptKT, and the Class Prior baseline. A PR curve reflects the trade-off between precision (the number of correct positive predictions) and recall (the model's ability to identify all positive instances). TIRT (AP = 0.7997) achieves the highest precision across all recall values, outperforming the other models. LR (AP = 0.7967) follows closely behind, and AdaptKT (AP = 0.7711) demonstrates the lowest precision of the top three models. The inset provides a zoomed-in view of the mid-recall region (between 0.52 and 0.68), offering a precise comparison of the closely aligned model curves and illustrating their relative performance.

Fig. 5(b) presents the ROC curves for the top three KT models, with the random curve (AUC = 0.5) as a baseline. The ROC curve illustrates each model's ability to distinguish between true and false positives. TIRT (AUC = 0.7468) demonstrates the greatest discriminatory ability, achieving the highest true positive rate at any given false positive rate. LR (AUC = 0.7446) follows closely behind, while AdaptKT (AUC = 0.7054) exhibits the weakest discriminatory ability. The inset zooms into the mid-range false positive rate region (between 0.40 and 0.60), providing a clearer perspective on the closely clustered model curves and highlighting TIRT's superior performance over LR and AdaptKT.

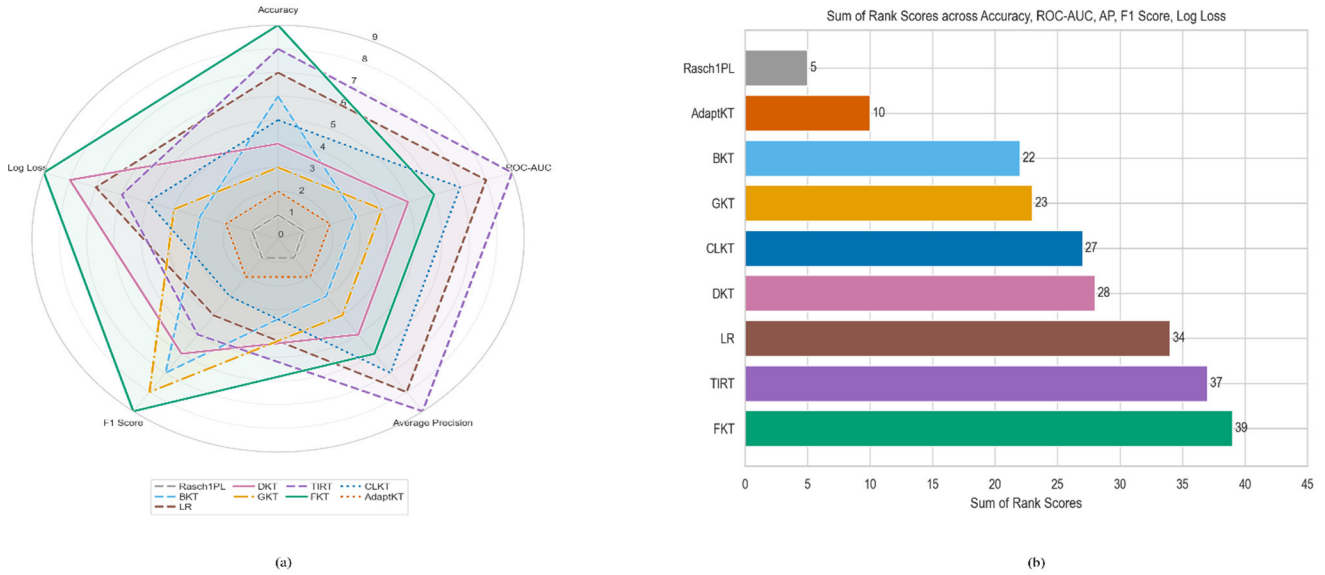

FIGURE 7. KT model performance and composite scores across DigiArvi 2025.

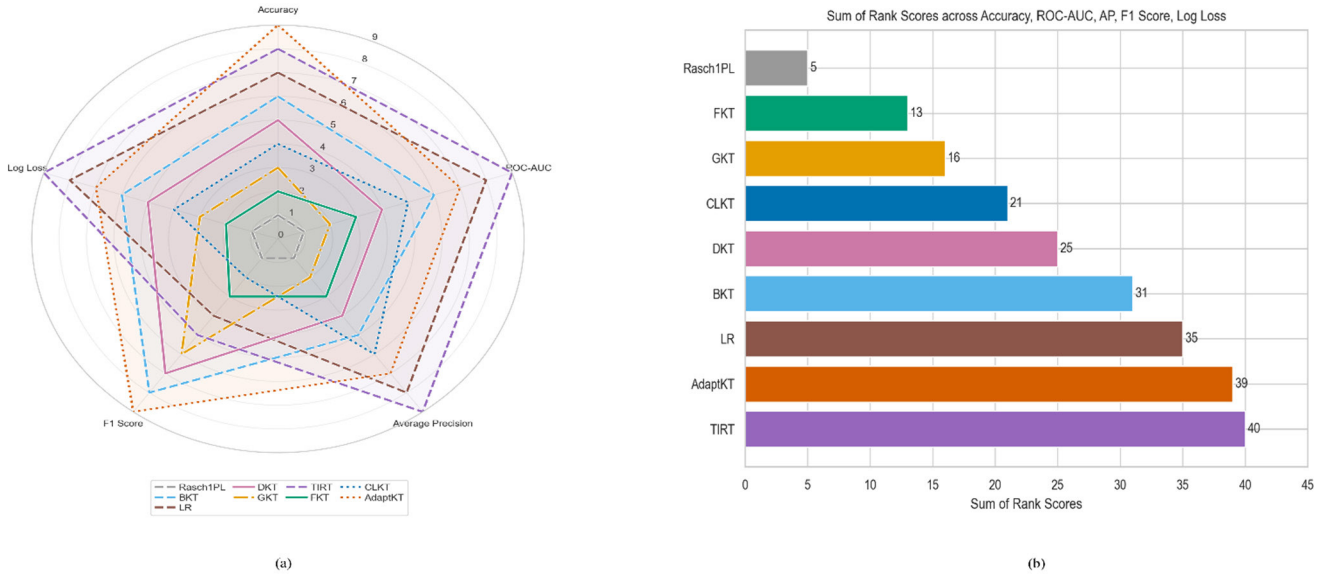

FIGURE 8. KT model performance and composite scores across ASSISTments 09-10.

Fig. 6(a) shows the PR curves for the top three KT models in DigiArvi: TIRT, LR, and CLKT, with the class prior baseline. TIRT leads among the models with the highest precision (AP = 0.8855), followed by LR (AP = 0.8847) and CLKT (AP = 0.8839). The inset zooms into the high precision, mid-recall region to provide a clearer comparison between the closely aligned curves.

Fig. 6(b) displays the ROC curves for the same models with the random baseline (AUC = 0.5). TIRT (AUC = 0.7991) demonstrates the greatest discriminatory power, followed closely by LR (AUC = 0.7981) and CLKT (AUC = 0.7971). The inset zooms into the mid-range false positive rate region, visually confirming TIRT's consistently superior performance.

In conclusion, comparing the performance of the top three KT models across the ASSISTments 09-10 and DigiArvi 2025 datasets shows that TIRT is effective in terms of precision and discriminatory power. TIRT consistently outperforms the other models, yielding the highest AP and AUC values in the PR and ROC curves, respectively. Although LR follows closely behind, TIRT's superior ability to balance precision and recall, along with its higher true positive rate, establishes it as the most robust model among those tested.

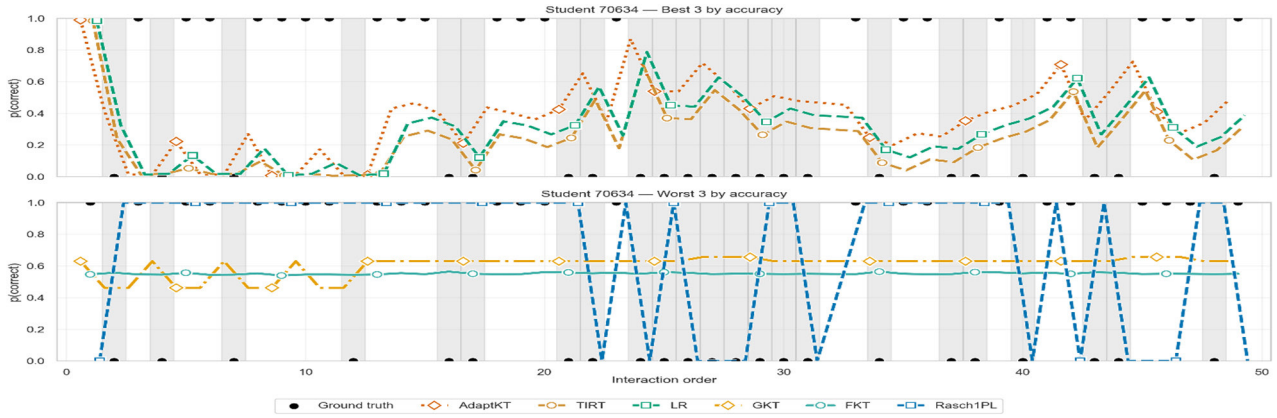

FIGURE 9. Performance of KT models in ASSISTments 09-10 on student 70634.

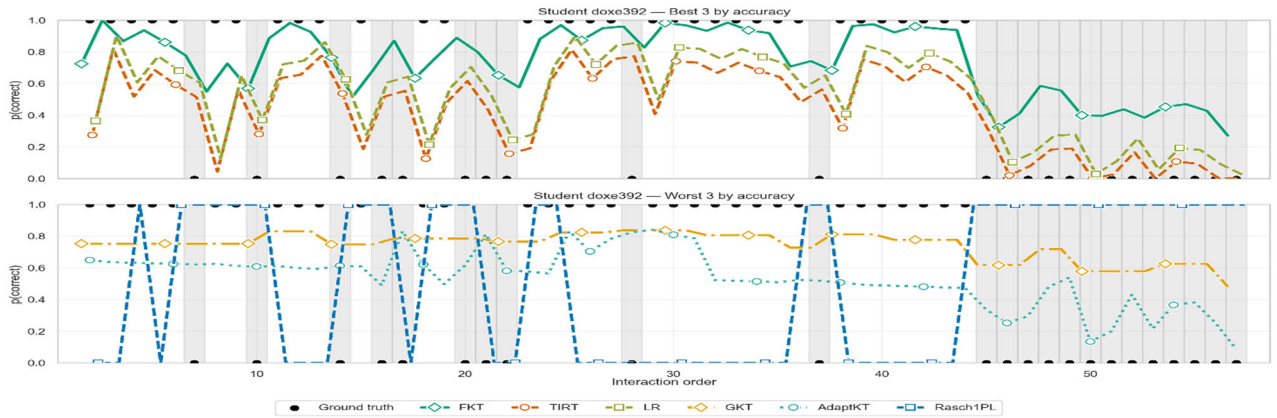

FIGURE 10. Performance of KT models in DigiArvi 2025 on student doxe392.

3) OVERALL MODEL PERFORMANCE RANKING

This section presents the overall performance ranking of various KT models using the ASSISTments 09-10 and DigiArvi 2025 datasets. The ranking is based on the following key evaluation metrics: accuracy, ROC-AUC, AP, F1 score, and log loss.

Fig. 7(a) is a radar chart visualizing the performance of different models across these five metrics in the DigiArvi 2025 dataset, based on the KT models' rank scores in each metric. FKT leads in overall performance, achieving consistently high rank scores across all metrics, especially in accuracy, F1 score, and log loss. TIRT and LR follow closely behind, demonstrating strong results in accuracy, ROC-AUC, and AP. Models such as DKT and CLKT demonstrate solid performance, though they fall short of the top three models.

Fig. 7(b) presents a bar chart summarizing the composite rank scores across the five evaluation metrics. The overall ranking aligns with the radar chart, with FKT leading by a significant margin. TIRT achieves the second-highest composite score, and LR achieves the third-highest, reflecting their competitive performance. In contrast, models such as AdaptKT and Rasch 1PL exhibit significantly lower scores, indicating comparatively weaker performance across the full spectrum of evaluation criteria.

These figures demonstrate the superiority of FKT and TIRT on the DigiArvi dataset. They appear to be the most promising models for predicting the probability that a student will answer the next item correctly in this context. The comparison highlights the variability in model performance across different evaluation criteria and emphasizes the importance of selecting the appropriate model based on specific requirements.

Fig. 8(a) is a radar chart that visualizes the performance of various models across five key metrics in the ASSISTments 09-10 dataset. TIRT demonstrates the strongest overall performance, excelling in AP, log loss, and ROC-AUC. AdaptKT follows in second place. LR shows competitive results, but falls behind TIRT and AdaptKT. In contrast, FKT and Rasch 1PL perform weakly across most metrics, with Rasch 1PL achieving the lowest scores.

Fig. 8(b) displays a bar chart summarizing the composite rank scores from the five evaluation metrics. TIRT leads with the highest total score, followed closely by AdaptKT. LR achieves the third-highest score, while FKT and Rasch

1PL significantly trail behind, highlighting their lower effectiveness across the evaluation criteria. These results reinforce the dominance of TIRT and AdaptKT in ASSISTments 09-10 and suggest that they are the most promising models for this dataset.

Although the overall composite scores of the models in the example datasets are clear, we observed notable fluctuations in the performance of AdaptKT and FKT. For example, AdaptKT achieved an overall rank score of 8 on ASSISTments 09-10 and 2 on DigiArvi 2025, while FKT achieved an overall rank score of 2 on ASSISTments 09-10 and 9 on DigiArvi 2025. These discrepancies are likely due to differences in dataset characteristics, such as response time variability, data consistency, and the models' varying sensitivities to noise and feature distributions. These performance variations suggest that models tend to perform better on datasets that align with their inherent architectural strengths.

4) PERFORMANCE OF KT MODELS ON INDIVIDUAL STUDENTS

This section evaluates the performance of various KT models on individual students from the ASSISTments 09-10 and DigiArvi 2025 datasets. The goal is to assess how well these models predict whether a student will answer the next question correctly based on their past interactions.

Fig. 9 and Fig. 10 compare the performance of different models for two students, one from the ASSISTments 09-10 dataset (ID: 70634) and one from the DigiArvi 2025 dataset (ID: doxe392). For each student, the top chart shows the three models with the highest accuracy. These models demonstrate strong alignment with the actual student performance, with predictive peaks and valleys that closely mirror the ground truth. Conversely, the bottom chart shows the three worst models, which fail to predict student performance accurately; their predictions diverge significantly from the actual outcomes. Specifically, for the student in the ASSISTments 09-10 dataset (Fig. 9), the best-performing models are AdaptKT, TIRT, and LR, while the worst are GKT, FKT, and Rasch 1PL. For the student in the DigiArvi 2025 dataset (Fig. 10), the best-performing models are FKT, TIRT, and LR, whereas the weakest performers are GKT, AdaptKT, and Rasch 1PL. These student-level observations directly mirror the overall composite rankings established in Section III-D3.

In conclusion, evaluating KT models across the ASSISTments 09-10 and DigiArvi 2025 datasets at the individual student level reveals significant differences in prediction accuracy. The best models closely track actual student learning trajectories, while the worst models show significant divergence. These findings visually underscore the importance of selecting models based on their ability to reliably predict student performance across diverse datasets.

IV. DISCUSSION

This study provides a thorough analysis of KT models across four key dimensions. First, we provided an overview of the ''big picture'' to illustrate the overlap and interconnections between the various KT model categories. Our findings reveal an increasing trend of integrating KT models with other domains to personalize the learning process, improve prediction accuracy, and adapt models to individual learners' needs. Second, we traced the timeline of KT model development, highlighting their evolution from foundational models in the 1950s to today's sophisticated deep learning techniques. Between 2015 and 2019, notable advances in deep learning, particularly in RNNs and LSTM networks, significantly influenced the design of modern KT models. These breakthroughs laid the groundwork for the most advanced models in use today, including multi-activity, domain-adaptive, and contrastive/self-supervised models. Third, we explored the different categories of KT models, such as psychometric, Bayesian, machine learning, deep learning, and graph-based models, demonstrating how each category contributes to developing personalized learning environments. For our empirical analysis, we selected a representative model from each category to demonstrate its practical application. Finally, we conducted a detailed empirical exploration of the selected models. We examined evaluation metrics, datasets, prediction curves, and student-level predictions to assess the models' effectiveness in two different datasets.

A. LIMITATIONS OF THE STUDY

Despite the rigorous methodology employed in this systematic review and empirical analysis, it is important to acknowledge several limitations to provide a balanced perspective on the findings. To ensure feasibility and strictly evaluate the efficiency of 'Green AI' [17], we selected lightweight representatives from each of the nine KT families for empirical evaluation. While these models are widely cited, they may not fully capture performance variance within their respective categories. For example, other graph-based models may outperform GKT [51], and different deep learning architectures may produce different results from DKT [40]. Consequently, our findings should be interpreted as an evaluation of efficient architectural families rather than as an exhaustive benchmark of every existing model. Furthermore, our empirical evaluation relied on two datasets: ASSISTments 09-10 [13] and DigiArvi 2025 [14]. Evaluating the models on just two datasets restricts the broader generalizability of our findings. Both datasets focus primarily on mathematics education. Therefore, the performance of these models may differ when applied to other domains, such as language learning, programming, or open-ended essay writing, where the nature of the knowledge components and student interactions differs significantly. Future benchmarking should incorporate datasets like Slepemapy, Junyi Academy, or KDD Cup to confirm these trends.

Additionally, while this study imposed a strict global training budget of 120 seconds per model to establish a baseline for algorithmic efficiency, a critical limitation is that our formal evaluation focused primarily on predictive metrics such as accuracy, AUC, and F1 score. It must be explicitly acknowledged that the predictive scores reported in this

study, particularly the ROC-AUC metrics, are lower than those often cited in state-of-the-art literature. This discrepancy is a direct and intentional consequence of our strict adherence to the 'Green AI' 120-second training budget. Instead of using heavily parameterized, resource-intensive models that take hours to train for only slight performance improvements, we deliberately evaluated lightweight baseline architectures to test true computational efficiency. A deeper micro-architectural analysis of computational costs was beyond the scope of this study. As recent literature has noted, 'vast and complex' deep learning models often incur high hidden costs in terms of continuous GPU utilization and energy consumption [17], [85]. As we did not benchmark precise inference latency (e.g., milliseconds per prediction) or exact energy footprints (e.g., joules consumed), we cannot definitively quantify the sustainability gains of the tested architectures. Therefore, while TIRT and FKT achieved high accuracy well within our 'Green AI' resource constraints, further empirical profiling of hardware utilization is required to determine the most computationally efficient solutions for real-time deployment. Finally, the systematic review adhered to the PRISMA 2020 guidelines [12] and was restricted to English-language studies published in Springer, ScienceDirect, Scopus, and Google Scholar between 2019 and 2025. These exclusion criteria may have omitted relevant studies published in non-indexed venues or in languages other than English, potentially introducing selection bias. Furthermore, due to the rapid pace of AI development, the 'state-of-the-art' cut-off point of 2025 may mean that newer architectures emerging during the publication process are not represented.

B. FUTURE RESEARCH DIRECTIONS AND PRACTICAL RECOMMENDATIONS

Based on our systematic review and rigorous empirical evaluation, we have identified critical challenges and provided actionable recommendations for the next generation of KT research. Recent trends suggest that the next probable evolution of KT is the convergence of numeric sequence modelling with generative AI. Current models rely on binary correctness data and therefore miss the semantic nuances of students' answers. Future algorithms must consequently integrate LLMs to process textual responses and generate automated pedagogical feedback [86]. The challenge lies in selecting the right architecture: researchers must determine how to fuse the high dimensionality of LLM embeddings with the temporal efficiency of traditional KT models without inducing excessive latency. To address these computational constraints, future research must pivot towards 'Green AI' by developing lightweight, distilled KT models that maintain high diagnostic accuracy while consuming less energy [17]. This would make them viable for real-time deployment in low-resource educational settings.

For practitioners currently deploying these models, our empirical assessment yields specific architectural recommendations. For sequential or temporal data, we recommend TIRT [84], as this model has proven to be the most robust for datasets with strong time-decay signals (such as ASSISTments), offering the best balance of accuracy and computational efficiency. For granular or multi-task data, we recommend FKT [68]. In modern environments that capture auxiliary data, such as response speeds in DigiArvi, FKT successfully leverages multi-task learning to outperform traditional baselines. Finally, if the priority is system diagnostics rather than raw prediction, LR remains a 'hard-to-beat' benchmark offering transparency that deep neural networks cannot yet match. To properly evaluate these and future advancements, the field urgently requires a unified benchmarking framework, similar to the General Language Understanding Evaluation (GLUE) in Natural Language Processing (NLP) [87], that evaluates models uniformly on accuracy, robustness, cold-start transferability, and inference speed. Ultimately, the next frontier for KT is moving from correlation to causation. Future architectures should leverage Multimodal Learning Analytics (MMLA) by combining log data with eye tracking or video, and apply causal inference to determine whether a student has truly mastered a skill or has simply guessed correctly [88]. This shift is essential for developing 'agentic AI' tutors that can effectively intervene rather than merely predict failure. Finally, as these advanced systems transition from controlled testing environments to real-world classrooms, researchers must address several ongoing challenges. Firstly, it is crucial to mitigate the 'cold-start' problem; models must be engineered to rapidly and accurately infer a baseline knowledge state for students with no historical interaction logs. Secondly, real-world educational data is inherently unreliable, necessitating more robust architectures that can filter out accidental errors, unrecorded offline learning, or system-level logging errors. Solving these practical challenges alongside the successful integration of rich, multimodal data streams will be vital for ensuring that future KT systems are highly accurate, scalable, and equitable for all learners.

V. CONCLUSION

This study provides a thorough, systematic review and rigorous empirical evaluation of KT models, mapping their development from foundational psychometric methods to modern deep learning paradigms. The research addressed critical gaps in our understanding of KT models by proposing a taxonomy of nine interconnected model families and by benchmarking representative algorithms across two educational datasets: ASSISTments (2009–2010) and DigiArvi (2025). Our empirical findings demonstrate that, while high-capacity deep learning models offer theoretical advantages in complex sequence modelling, their practical application often yields diminishing returns. Instead, architectures that are strategically engineered to leverage specific data characteristics—such as the explicit modelling of forgetting in TIRT and the integration of auxiliary features in multi-task FKT—consistently achieve superior predictive calibration and student-level diagnostic accuracy, even when restricted by strict computational time budgets. Furthermore, traditional

baselines such as LR remain competitive, offering a level of diagnostic transparency that deep neural networks cannot yet match.

This study goes beyond raw predictive performance, highlighting the critical engineering trade-offs necessary for the real-world, scalable deployment of intelligent tutoring systems. The success of these lightweight yet highly robust models highlights the urgent need for the EDM community to adopt 'Green AI' principles. This will ensure that future architectures are computationally sustainable, economically viable, and accessible in low-resource educational environments. Furthermore, the fragmented evaluation criteria currently prevalent in the field emphasize the necessity of establishing standardized, open-source benchmarking frameworks to guarantee reproducibility. Looking to the future, the next frontier of knowledge tracing is integrating LLMs and causal inference mechanisms to transform KT frameworks into agentic AI systems that can deliver autonomous, active, and deeply personalized pedagogical interventions.

DATA AVAILABILITY

The DigiArvi 2025 dataset used in this study is available upon request for reviewers from the corresponding author. The ASSISTments 09-10 dataset can be accessed at https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data

ETHICAL APPROVAL AND INFORMED CONSENT

The data used in this study were obtained from two sources: DigiArvi 2025 and ASSISTments 09-10. While the DigiArvi 2025 dataset is not publicly accessible, it has been anonymized. The ASSISTments 09-10 dataset is publicly available and does not contain any personally identifiable information. Under Finnish law, informed consent is not required when the number of participants exceeds 400. Instead, participants were allowed to opt out of the study if they did not wish to participate.

ACKNOWLEDGMENT

During the preparation of this manuscript, the authors used DeepL, Grammarly, and Gemini 3 Pro to improve spelling, grammar, and clarity of expression. All content generated with the help of these tools was carefully checked and edited to ensure it was accurate and original. They also take full responsibility for the final version of the manuscript.

REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge tracing: A survey," *ACM Comput. Surveys*, vol. 55, no. 11, pp. 1–37, Nov. 2023, doi: 10.1145/3569576.

[2] S. Shen, Q. Liu, Z. Huang, Y. Zheng, M. Yin, M. Wang, and E. Chen, "A survey of knowledge tracing: Models, variants, and applications," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 1858–1879, Apr. 2024, doi: 10.1109/TLT.2024.3383325.

[3] E. H. Am, I. Hidayah, and S. S. Kusumawardani, "A literature review of knowledge tracing for student modeling: Research trends, models, datasets, and challenges," *J. Inf. Technol. Comput. Sci.*, vol. 6, no. 2, pp. 183–194, Oct. 2021, doi: 10.25126/jitecs.202162344.

[4] X. Sun, X. Zhao, Y. Ma, X. Yuan, F. He, and J. Feng, "Muti-behavior features based knowledge tracking using decision tree improved DKVMN," in *Proc. ACM Turing Celebration Conf. China*. Chengdu, China: ACM, May 2019, pp. 1–6, doi: 10.1145/3321408.3322847.

[5] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing AutoML in educational data mining for prediction tasks," *Appl. Sci.*, vol. 10, no. 1, p. 90, Dec. 2019, doi: 10.3390/app10010090.

[6] Y. Badran and C. Preisach, "Sparse binary representation learning for knowledge tracing," 2025, *arXiv:2501.09893*.

[7] W. Yukun, X. Xingjian, and M. Fanjun, "Deep knowledge tracking model integrating multiple feature personalization factors," in *Proc. IEEE Cyber Sci. Technol. Congr. (CyberSciTech)*, Nov. 2024, pp. 394–399, doi: 10.1109/cyberscitech64112.2024.00068.

[8] S. An, J. Kim, M. Kim, and J. Park, "No task left behind: Multitask learning of knowledge tracing and option tracing for better student assessment," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 4424–4431, doi: 10.1609/aaai.v36i4.20364.

[9] S. Cheng, Q. Liu, E. Chen, K. Zhang, Z. Huang, Y. Yin, X. Huang, and Y. Su, "AdaptKT: A domain adaptable method for knowledge tracing," in *Proc. 15th ACM Int. Conf. Web Search Data Mining*, Feb. 2022, pp. 123–131, doi: 10.1145/3488560.3498379.

[10] O. E. Aissaoui, L. Oughdir, and Y. E. Allioui, "A literature review on student modeling purposes," in *Proc. Adv. Intell. Syst. Comput.*, vol. 1417, 2022, pp. 758–784, doi: 10.1007/978-3-030-90633-7_64.

[11] A. Robitzsch, "Regularized generalized logistic item response model," *Information*, vol. 14, no. 6, p. 306, May 2023, doi: 10.3390/info14060306.

[12] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. 71, Mar. 2021, doi: 10.1136/bmj.n71.

[13] ASSISTment. (2015). *2009-2010 Assistment Data*. Google. [Online]. Available: https://sites.google.com/site/assistmentsdata/home/2009-2010-assistment-data

[14] J. Metsämuuronen and M.-J. Laakso, "DigiEva 2025 technical report ver 1.3," Turku Res. Inst. Learn. Anal. (TRILA), Univ. Turku, 2025, doi: 10.13140/RG.2.2.34370.03529.

[15] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, pp. 1–14, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[16] A. Aggarwal, S. Prasad Kasiviswanathan, Z. Xu, O. Feyisetan, and N. Teissier, "Label inference attacks from log-loss scores," 2021, *arXiv:2105.08266*.

[17] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Nov. 2020, doi: 10.1145/3381831.

[18] G. Rasch, *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Nielsen & Lydiche, 1960.

[19] F. M. Lord, M. R. Novick, and A. Birnbaum, *Statistical Theories of Mental Test Scores*. Reading, MA, USA: Addison-Wesley, 1968.

[20] R. D. Bock and M. Aitkin, "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, vol. 46, no. 4, pp. 443–459, Dec. 1981, doi: 10.1007/bf02293801.

[21] D. R. Cox, "The regression analysis of binary sequences," *J. Roy. Stat. Soc. Ser. B, Stat. Methodology*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[22] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *J. ACM*, vol. 7, no. 3, pp. 216–244, Jul. 1960, doi: 10.1145/321033.321035.

[23] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1023/a:1022643204877.

[24] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artif. Intell.*, vol. 29, no. 3, pp. 241–288, Sep. 1986, doi: 10.1016/0004-3702(86)90072-x.

[25] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297. [Online]. Available: https://www.cs.cmu.edu/ bhiksha/courses/mlsp.fall2010/class14/macqueen.pdf

[26] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapted Interact.*, vol. 4, no. 4, pp. 253–278, 1995, doi: 10.1007/bf01099821.

[27] R. S. J. d Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing," in *Proc. Intell. Tutoring Syst.*, vol. 5091, 2008, pp. 406–415, doi: 10.1007/978-3-540-69132-7_44.

[28] K. Chang, J. Beck, J. Mostow, and A. Corbett, "A Bayes net toolkit for student modeling in intelligent tutoring systems," in *Intelligent Tutoring Systems* (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2006, pp. 104–113, doi: 10.1007/11774303_11.

[29] J. Reye, "Student modelling based on belief networks," *Int. J. Artif. Intell. Educ.*, vol. 14, no. 1, pp. 63–96, Feb. 2004, doi: 10.3233/irg-2004-14(1)04.

[30] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis—A general method for cognitive model evaluation and improvement," in *Intelligent Tutoring Systems* (Lecture Notes in Computer Science), vol. 4053. Berlin, Germany: Springer, 2006, pp. 164–175, doi: 10.1007/11774303_17.

[31] P. I. Pavlik Jr., H. Cen, and K. R. Koedinger, "Performance factors analysis—A new alternative to knowledge tracing," in *Proc. 14th Int. Conf. Artif. Intell. Educ.*, 2009, pp. 1–8. [Online]. Available: https://files.eric.ed.gov/fulltext/ED506305.pdf

[32] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/bf00994018.

[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.

[34] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, Dec. 1992, doi: 10.1145/138859.138867.

[35] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1–8. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2007/file/d7322ed717dedf1eb4e6e52a37ea7bcd-Paper.pdf

[36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., Cambridge, MA, USA: MIT Press, 2014. [Online]. Available: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

[37] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, May 1992, doi: 10.1007/bf00992698.

[38] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a Bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization* (Lecture Notes in Computer Science), vol. 6075. Berlin, Germany: Springer, 2010, pp. 255–266, doi: 10.1007/978-3-642-13470-8_24.

[39] Y. Wang and N. Heffernan, "Towards modeling forgetting and relearning in ITS: Preliminary analysis of ARRS data," in *Proc. 4th Int. Conf. Educ. Data Mining*, 2011, pp. 1–6. [Online]. Available: https://web.cs.wpi.edu/ nth/pubs_and_grants/papers/2011/EDM%202011/Wang%20Towards%20Modeling%20Forgetting%20and%20Relearning.pdf

[40] C. Piech, J. Bassen, and J. Huang, "Deep knowledge tracing," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf

[41] L. Zhang, X. Xiong, S. Zhao, A. Botelho, and N. T. Heffernan, "Incorporating rich features into deep knowledge tracing," in *Proc. 4th ACM Conf. Learn. Scale*, Apr. 2017, pp. 169–172, doi: 10.1145/3051457.3053976.

[42] K. Nagatani, Q. Zhang, M. Sato, Y.-Y. Chen, F. Chen, and T. Ohkuma, "Augmenting knowledge tracing by considering forgetting behavior," in *Proc. World Wide Web Conf.*, 2019, pp. 3101–3107, doi: 10.1145/3308558.3313565.

[43] G. K. Dziugaite and D. M. Roy, "Neural network matrix factorization," 2015, *arXiv:1511.06443*.

[44] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C. Ding, S. Wei, and G. Hu, "Exercise-enhanced sequential modeling for student performance prediction," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 2435–2443, doi: 10.1609/aaai.v32i1.11864.

[45] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu, "EKT: Exercise-aware knowledge tracing for student performance prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 1, pp. 100–115, Jan. 2021, doi: 10.1109/TKDE.2019.2924374.

[46] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 765–774, doi: 10.1145/3038912.3052580.

[47] G. Abdelrahman and Q. Wang, "Knowledge tracing with sequential key-value memory networks," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 175–184, doi: 10.1145/3331184.3331195.

[48] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937. [Online]. Available: https://proceedings.mlr.press/v48/mniha16.pdf

[49] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[50] D. Cai, Y. Zhang, and B. Dai, "Learning path recommendation based on knowledge tracing model and reinforcement learning," in *Proc. IEEE 5th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2019, pp. 1881–1885, doi: 10.1109/ICCC47050.2019.9064104.

[51] H. Nakagawa, Y. Iwasawa, and Y. Matsuo, "Graph-based knowledge tracing: Modeling student proficiency using graph neural network," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Oct. 2019, pp. 156–163, doi: 10.1145/3350546.3352513.

[52] C.-K. Yeung, "Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory," 2019, *arXiv:1904.11738*.

[53] Y. Choi, Y. Lee, J. Cho, J. Baek, B. Kim, Y. Cha, D. Shin, C. Bae, and J. Heo, "Towards an appropriate query, key, and value computation for knowledge tracing," in *Proc. 7th ACM Conf. Learn. Scale*, Aug. 2020, pp. 341–344, doi: 10.1145/3386527.3405945.

[54] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi, "SAINT+: Integrating temporal features for EdNet correctness prediction," in *Proc. LAK21: 11th Int. Learn. Analytics Knowl. Conf.*, Apr. 2021, pp. 490–496, doi: 10.1145/3448139.3448188.

[55] A. Ghosh, N. Heffernan, and A. S. Lan, "Context-aware attentive knowledge tracing," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2330–2339, doi: 10.1145/3394486.3403282.

[56] D. Sheng, "Grasping or forgetting? MAKT: A dynamic model via multi-head self-attention for knowledge tracing," in *Proc. Int. Conferences Softw. Eng. Knowl. Eng.*, vol. 2021, Jul. 2021, pp. 399–404, doi: 10.18293/seke2021-031.

[57] L. Li and Z. Wang, "Calibrated Q-matrix-enhanced deep knowledge tracing with relational attention mechanism," *Appl. Sci.*, vol. 13, no. 4, p. 2541, Feb. 2023, doi: 10.3390/app13042541.

[58] D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang, and Y. Yin, "Multiple features fusion attention mechanism enhanced deep knowledge tracing for student performance prediction," *IEEE Access*, vol. 8, pp. 194894–194903, 2020, doi: 10.1109/ACCESS.2020.3033200.

[59] X. Song, J. Li, Y. Tang, T. Zhao, Y. Chen, and Z. Guan, "JKT: A joint graph convolutional network based deep knowledge tracing," *Inf. Sci.*, vol. 580, pp. 510–523, Nov. 2021, doi: 10.1016/j.ins.2021.08.100.

[60] H. Tong, Z. Wang, Y. Zhou, S. Tong, W. Han, and Q. Liu, "HGKT: Introducing hierarchical exercise graph for knowledge tracing," 2020, *arXiv:2006.16915*.

[61] Z. Wu, L. Huang, Q. Huang, C. Huang, and Y. Tang, "SGKT: Session graph-based knowledge tracing for student performance prediction," *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117681, doi: 10.1016/j.eswa.2022.117681.

[62] G. Abdelrahman and Q. Wang, "Deep graph memory networks for forgetting-robust knowledge tracing," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 1–13, 2022, doi: 10.1109/TKDE.2022.3206447.

[63] L. Wei, B. Li, Y. Li, and Y. Zhu, "Time interval aware self-attention approach for knowledge tracing," *Comput. Electr. Eng.*, vol. 102, Sep. 2022, Art. no. 108179, doi: 10.1016/j.compeleceng.2022.108179.

[64] C. Wang, W. Ma, M. Zhang, C. Lv, F. Wan, H. Lin, T. Tang, Y. Liu, and S. Ma, "Temporal cross-effects in knowledge tracing," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 517–525, doi: 10.1145/3437963.3441802.

[65] S. Zu, L. Li, and J. Shen, "CAKT: Coupling contrastive learning with attention networks for interpretable knowledge tracing," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–8, doi: 10.1109/ijcnn54540.2023.10191799.

[66] W. Lee, J. Chun, Y. Lee, K. Park, and S. Park, "Contrastive learning for knowledge tracing," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2330–2338, doi: 10.1145/3485447.3512105.

[67] T. Wu and Q. Ling, "Self-supervised heterogeneous hypergraph network for knowledge tracing," *Inf. Sci.*, vol. 624, pp. 200–216, May 2023, doi: 10.1016/j.ins.2022.12.075.

[68] T. Huang, S. Hu, H. Yang, J. Geng, Z. Li, Z. Xu, and X. Ou, "Response speed enhanced fine-grained knowledge tracing: A multi-task learning perspective," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 122107, doi: 10.1016/j.eswa.2023.122107.

[69] X.-L. Diao, C.-H. Zheng, Q.-T. Zeng, H. Duan, Z.-G. Song, and H. Zhao, "Precise modeling of learning process based on multiple behavioral features for knowledge tracing," *J. Intell. Fuzzy Syst.*, vol. 44, no. 6, pp. 10747–10764, Jun. 2023, doi: 10.3233/jifs-224351.
[70] S. Zhao, C. Wang, and S. Sahebi, "Transition-aware multi-activity knowledge tracing," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2022, pp. 1760–1769, doi: 10.1109/BigData55660.2022.10020617.
[71] C.-Q. Huang, Q.-H. Huang, X. Huang, H. Wang, M. Li, K.-J. Lin, and Y. Chang, "XKT: Toward explainable knowledge tracing model with cognitive learning theories for questions of multiple knowledge concepts," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 7308–7325, Nov. 2024, doi: 10.1109/TKDE.2024.3418098.
[72] S. Minn, J.-J. Vie, K. Takeuchi, H. Kashima, and F. Zhu, "Interpretable knowledge tracing: Simple and efficient student modeling with causal relations," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 12810–12818, doi: 10.1609/aaai.v36i11.21560.
[73] Y. Su, Z. Cheng, P. Luo, J. Wu, L. Zhang, Q. Liu, and S. Wang, "Time-and-concept enhanced deep multidimensional item response theory for interpretable knowledge tracing," *Knowledge-Based Syst.*, vol. 218, Apr. 2021, Art. no. 106819, doi: 10.1016/j.knosys.2021.106819.
[74] Y. Bai, J. Zhao, T. Wei, Q. Cai, and L. He, "A survey of explainable knowledge tracing," *Appl. Intell.*, vol. 54, no. 8, pp. 6483–6514, Apr. 2024, doi: 10.1007/s10489-024-05509-8.
[75] P. I. S. Lei and A. J. Mendes, "A systematic literature review on knowledge tracing in learning programming," in *Proc. IEEE Frontiers Educ. Conf. (FIE)*, Oct. 2021, pp. 1–7, doi: 10.1109/FIE49875.2021.9637323.
[76] I. Šarić-Grgić, A. Grubišić, and A. Gašpar, "Twenty-five years of Bayesian knowledge tracing: A systematic review," *User Model. User-Adapted Interact.*, vol. 34, no. 4, pp. 1127–1173, Sep. 2024, doi: 10.1007/s11257-023-09389-4.
[77] H. Zhou, R. Bamler, C. M. Wu, and Á. Tejero-Cantero, "Predictive, scalable and interpretable knowledge tracing on structured domains," 2024, *arXiv:2403.13179*.
[78] B. Jiang, S. Wu, C. Yin, and H. Zhang, "Knowledge tracing within single programming practice using problem-solving process data," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 822–832, Oct. 2020, doi: 10.1109/TLT.2020.3032980.
[79] S. I. R. Luelmo, N. E. Mawas, and J. Heutte, "Existing machine learning techniques for knowledge tracing: A review using the PRISMA guidelines," in *Proc. Commun. Comput. Inf. Sci.*, vol. 1624, 2022, pp. 73–94, doi: 10.1007/978-3-031-14756-2_5.
[80] M. U. Khan, S. Mehak, D. W. Yasir, S. Anwar, M. U. Majeed, and H. A. Ramzan, "Quantitative studies of deep reinforcement learning in gaming, robotics and real-world control systems," *Bull. Bus. Econ. (BBE)*, vol. 12, no. 2, pp. 389–395, Aug. 2023, doi: 10.61506/01.00019.
[81] S. Zheng, Q. Xiong, Y. Li, T. Han, and J. Guo, "Innovating educational assessment: A hybrid TCN-LSTM model for knowledge tracing," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2024, pp. 3643–3649, doi: 10.1109/smc54092.2024.10832002.
[82] Y. Lai, X. Xu, X. Zhang, X. Dong, J. Zhuang, and J. Liu, "Improving knowledge tracing through learning processes and concept similarity map," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2024, pp. 4840–4847, doi: 10.1109/smc54092.2024.10831394.
[83] L. Qian, K. Zheng, L. Wang, and S. Li, "Student state-aware knowledge tracing based on attention mechanism: A cognitive theory view," *Pattern Recognit. Lett.*, vol. 184, pp. 190–196, Aug. 2024, doi: 10.1016/j.patrec.2024.06.009.
[84] J. González-Brenes, Y. Huang, and P. Brusilovsky, "General features in knowledge tracing: Applications to multiple subskills, temporal item response theory, and expert knowledge," in *Proc. 7th Int. Conf. Educ. Data Mining*, 2014, pp. 84–91. Accessed: Feb. 18, 2026. [Online]. Available: https://d-scholarship.pitt.edu/26017
[85] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," 2019, *arXiv:1906.02243*.
[86] Y. Cho, R. E. AlMamlook, and T. Gharaibeh, "A systematic review of knowledge tracing and large language models in education: Opportunities, issues, and future research," 2024, *arXiv:2412.09248*.
[87] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2018, *arXiv:1804.07461*.
[88] R. Martinez-Maldonado, V. Echeverria, G. Fernandez-Nieto, L. Yan, L. Zhao, R. Alfredo, X. Li, S. Dix, H. Jaggard, R. Wotherspoon, A. Osborne, D. Gašević, and S. B. Shum, "Lessons learnt from a multimodal learning analytics deployment in-the-wild," 2023, *arXiv:2303.09099*.

PRINCE DAS ADHIKARY received the bachelor's degree in computer technology and the M.Sc. degree in finance and big data analytics from Swansea University, U.K. He is currently pursuing the Ph.D. degree in artificial intelligence-driven education and learning analytics with the University of Turku, Finland. He is also a Ph.D. Researcher with Turku Research Institute for Learning Analytics (TRILA), University of Turku. Before his academic research, he was a Software Engineer with more than three years of experience specializing in scalable software solutions and Java development. His primary research interests include knowledge tracing, machine learning, adaptive learning systems, and educational data mining.

JARI METSÄMUURONEN received the M.A. degree in education, in 1991, the degree in statistics, in 1993, and the Ph.D. degree in education in 1995. He is currently a Professor of learning analytics and the Head of the Analytic Team, Turku Research Institute for Learning Analytics (TRILA), University of Turku, Finland. Before this role, he was a Counsellor of evaluation with Finnish Education Evaluation Centre (FINEEC) and a Professor of education with the NLA University College, Norway. Outside of academia, he has been running a developmental berry-farming project to support local farmers in Nepal, since 2013. His research interests include psychometrics, evaluation research, research methodology, and the development of next-generation psychometric tools for digital learning environments.

MIKKO-JUSSI LAAKSO is currently a Professor and the Director of Turku Research Institute for Learning Analytics (TRILA), University of Turku. He specializes in learning analytics, digitally supported learning, pedagogical systems, programming education, gamification, and machine learning in education. He has played a key role in the development of the ViLLE learning systems, a digitally supported learning environment widely used in Finnish schools that has been recognized with a UNESCO Award. In addition, he is the Deputy Director of the EDUCA Flagship Network, University of Turku. Outside of his academic career, he serves as a top-level referee for Superpesis, the premier league of Finnish baseball. He has published numerous peer-reviewed scientific articles on learning analytics and has extensive experience in the research and development of user interfaces, assessment systems, and digital pedagogy.

JUKKA HEIKKONEN received the M.Sc. (Eng.) and Dr.Tech. degrees from Lappeenranta University of Technology, Finland, in 1991 and 1994, respectively. He is currently a Professor of computer science and the Head of the Data Analytics Research Group, University of Turku, Finland. Previously, he was a Scientific Officer for European Commission at the Joint Research Centre (JRC), Ispra, Italy, and a Senior Research Scientist with Helsinki University of Technology, Finland. He holds three patents in Finland. His research interests include artificial intelligence, data analytics, machine learning, and computational modeling. He was a recipient of the Award of Finnish Engineering Society for Innovative Engineering, in 1991.